10-28-2019

Deciphering the Gene Expression Control in Epigenetic, Post-transcriptional and Translational Regulation

Jianan Lin
University of Connecticut - Storrs, jianan.lin@uconn.edu

Deciphering the Gene Expression Control in Epigenetic, Post-transcriptional and Translational Regulation

Jianan Lin, PhD

University of Connecticut, 2019

Regulation of gene expression, through which cells increase or decrease the gene products, is an essential part of development, which not only determines the cellular differentiation but also responds to the environmental changes. A wide range of gene regulation mechanisms are involved in two major processes, from Deoxyribonucleic acid (DNA) to Ribonucleic acid (RNA), and from RNA to protein, also known as transcription and translation, respectively. Though the regulation of gene expression is not fully understood, this complex process has been characterized to several major steps, which includes epigenetic regulation, transcriptional regulation, post-transcriptional regulation and translational regulation. The present study covers three main projects related to three of the aforementioned steps of gene regulation. First, we study the DNA methylation dynamics in the bovine early embryos. To understand the epigenetic reprogramming and regulation in the embryonic development, we characterize the methylation process at the single-base level in early embryos. Second, we develop a novel method, called Protein-RNA Association Strength (PRAS), to predict the functional targets of RNA-Binding Proteins (RBPs) that play important roles in the regulation of gene expression in the post-transcriptional process (Keene 2007). The development of various Cross-linking and immunoprecipitation with high-throughput sequencing (CLIP-seq) data makes it possible to investigate the transcriptomic binding sites of RBPs (Licatalosi et al. 2008; Hafner et al. 2010; Konig et al. 2010; Konig et al. 2012; Cook et al. 2015). We aim to fill the gap between the peak-calling methods and the interpretation of RBPs' biological functions based on CLIP-seq data. Third, we study the regulation of c-MYC on the mRNA translation. The oncogenic c-MYC (MYC) transcription factor has broad effects on gene expression and cell behavior. We study how MYC affects the global translation of mRNAs and the translation start-site usage in the human lymphoma cell line. In sum, we perform data analysis and methodology development in these three specific projects related to the regulation of gene expression, which will help us better understand the central dogma of biology.

Deciphering the Gene Expression Control in Epigenetic, Post-transcriptional and Translational Regulation

Jianan Lin

B.S., Tianjin University, **2013**

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Connecticut

2019

APPROVAL PAGE

Doctor of Philosophy Dissertation

Deciphering the Gene Expression Control in Epigenetic, Post-transcriptional and Translational Regulation

Presented by

Jianan Lin, B.S.

Major Advisor _____
Zhengqing Ouyang

Associate Advisor _____
Yuping Zhang

Associate Advisor _____
Ion Mandoiu

University of Connecticut
2019

ACKNOWLEDGEMENTS

I dedicate this thesis to you, my dear grandpa. Even though you are no longer with us, I know you will be happy for me in heaven. Without your encouragement, I would not have taken the step to go abroad. I hope you are proud of your grandson, Dr. Jianan Lin.

First, I would like to thank my supervisor, Dr. Zhengqing Ouyang, for his expertise, patience, guidance and kindness throughout the entire process of my Doctor of Philosophy degree. Without your help, my works in research would not have been possible. I would like to thank my committee members, Dr. Yuping Zhang and Dr. Ion Mandoiu, for their suggestions, guidance and encouragement.

I would like to extend my gratitude to my excellent collaborators. I would like to give special thanks to Dr. Haifan Lin, Dr. Xiekui Cui, Dr. Hongying Qi, Dr. Hans-Guido Wendel, Dr. Kamini Singh, Dr. Xiuchun Cindy Tian, and Dr. Zongliang Jiang. Thank you all for your guidance and suggestions in the collaboration projects. I would also like to thank all my lab mates, including Dr. Yizhou Li, Dr. Chenchen Zou, Dr. Yubing Wan, Dr. Deepak Nag Ayyala, Dr. Yang Chen, Dr. Disheng Mao, and Youzhi Anthony Cheng, for your helpful suggestions as a lab member.

I would also like to thank my friends, Zichao Bian, Xiaoyu Ma, Xinhua Wu, Xin Zhou, Cheng-han Chung, Hao He, Yuqi Zhao, Mingze Sun, Kaikai Guo, and Jun Liao for giving me suggestions on the long journey.

Last but not least, I would like to thank my family members and give special thanks to my love, Anshuang. I am really lucky to have a family with you. All your support and sweetness is the foundation of my pursuing the PhD degree. You define my future.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1 GENE EXPRESSION REGULATION

Gene expression has two major steps: transcription which is from Deoxyribonucleic acid (DNA) to Ribonucleic acid (RNA) and translation which is from RNA to protein. A wide range of regulation mechanisms are involved in these two major processes, which include epigenetic regulation, transcriptional regulation, post-transcriptional regulation and translational regulation. In this thesis, we will cover a project related to DNA methylation because DNA methylation provides a stable, heritable, and critical component of epigenetic regulation (Goldberg et al. 2007). We will also include one study about RNA binding proteins (RBPs). RBPs play critical roles in RNAs' biogenesis, stability, function, transport and cellular localization (Glisovic et al. 2008). We will cover a project related to translational regulation in cancer cell line based on the fact that the translational regulation of oncogene expression is implicated in many cancers (Wolfe et al. 2014). Taken together, three aforementioned major processes of the gene expression control are studied in this thesis, including epigenetic, post-transcriptional and translational regulation. We will describe the concept of each of them in details in the following sections.

1.1.1 DNA METHYLATION

As mentioned, DNA methylation is an essential component of epigenetic regulation. DNA methylation is a process by which methyl groups are added to the DNA molecule. Methylation can change the activity of a DNA segment without changing the sequence. In mammals, nearly all DNA methylation is cytosine methylation which occurs on CG dinucleotides, with the cytosines on both strands being methylated. Regions of the genome that have a high density of CpGs are referred to as CpG islands, and DNA methylation of these islands correlates with transcriptional repression (Goll and Bestor 2005). This epigenetic process regulates gene transcription for differentiation, gene imprinting, and X-chromosome inactivation (Li et al. 1993; Bird 2002; Jaenisch and Bird 2003). Dynamic epigenetic modification of the genome occurs during early development of the mammals. The most dramatic genome-wide methylation changes occur in primordial germ cells and during preimplantation development (Smallwood et al. 2011;

Seisenberger et al. 2012; Smith et al. 2012; Guo et al. 2014; Smith et al. 2014; Gkountela et al. 2015; Guo et al. 2015; Gao et al. 2017). In early embryos, this involves the ultimate removal of cytosine methylation acquired in the gametes prior to fertilization, a process extensively characterized at the single-base level in the mouse, and more recently in humans (Smallwood et al. 2011; Smith et al. 2012; Guo et al. 2014; Smith et al. 2014).

1.1.2 RNA BINDING PROTEIN

RBPs are proteins that bind to the double or single stranded RNAs and form ribonucleoprotein complexes. The RBPs influence the structure and interactions of the RNAs and play critical roles in their biogenesis, stability, function, transport and cellular localization (Glisovic et al. 2008). Eukaryotic cells encode a large number of RBPs, each of which has unique RNA-binding activity characteristics and roles in the regulation of post-transcriptional gene expression (Glisovic et al. 2008). Technologies that comprehensively identify the RNA-RBP interactions are developed to discover the mechanisms by which RBPs affect RNA processing, such as RNA immunoprecipitation (RIP) and Cross-linking and immunoprecipitation (CLIP) (Ule et al. 2003; Gilbert et al. 2004).

1.1.3 RNA TRANSLATION

RNA translation is the second step of the central dogma of molecular biology, which encodes the protein based on the genetic information in RNA. Protein synthesis is regulated mainly at the translation initiation step, which controls the gene expression (Jackson et al. 2010). In this translational control of gene expression, a wide range of regulatory elements have been detected, which leads to a greater understanding of the translational control mechanisms (Wilkie et al. 2003). In the process of cell growth and development, the loss of certain translational regulation can contribute to the initiation and progression of cancer. Therefore, altering the protein synthesis machinery is one of the best-known functions of certain tumor suppressors and oncogenes (Ruggero and Pandolfi 2003).

1.2 RELATED HIGH-THROUGHPUT SEQUENCING TECHNOLOGIES

Reduced representation bisulfite sequencing (RRBS) was originally designed to analyze and compare genomic methylation patterns as a large-scale approach (Meissner et al. 2005). The principle of this technology is that bisulfite converts unmethylated cytosines to uracil (Meissner et al. 2005). By dividing the number of reported C ("methylated" reads) by the sum of reported C ("methylated" reads) and T ("unmethylated" reads) at the same positions of the reference genome, the methylation levels of each sampled cytosine can be estimated.

CLIP-seq was originally developed as a method to extract protein-RNA complexes from mouse brain with the use of ultraviolet cross-linking and immunoprecipitation (Ule et al. 2003). Improved versions of CLIP approaches have been developed and widely used to detect the binding peaks of RBPs at the transcriptome scale (Licatalosi et al. 2008; Hafner et al. 2010; Konig et al. 2010; Konig et al. 2012; Cook et al. 2015).

Ribo-seq, as known as ribosome profiling, is based on the deep sequencing of ribosome-protected mRNA fragments, which investigates the status of mRNA translation (Ingolia et al. 2009; Ingolia et al. 2011). Ribo-seq can be used to monitor the mRNA translation efficiency by correcting with the mRNA abundance. In detail, the translational efficiency is calculated from the ratio of ribosome footprint density from ribo-seq data to mRNA abundance from mRNA-seq data (Ingolia et al. 2011).

RNA-seq, one of the most widely used sequencing technology, was developed to quantify the eukaryotic transcriptomes (Cloonan et al. 2008; Lister et al. 2008; Marioni et al. 2008; Morin et al. 2008; Mortazavi et al. 2008; Nagalakshmi et al. 2008; Wilhelm et al. 2008; Wang et al. 2009). Since the read count bias is commonly presented in the RNA-seq data, normalization is a key step in analyzing it. There are mainly two sources of the bias, the sequencing depth and the gene length (Finotello and Di Camillo 2015). Global scaling, defined as scaling each gene's read count by the total number of mapped reads, was originally used for the purpose of addressing the sequencing depth (Marioni et al. 2008; Mortazavi et al. 2008; Finotello and Di Camillo 2015). However, recently normalization methods of sequencing depth have been developed to correcting the over-representation of high-expressed genes (Anders and Huber 2010;

Robinson and Oshlack 2010; Lin et al. 2011; Li et al. 2012; Love et al. 2014). As for the gene length normalization, reads or fragments per kilobase of exon (RPK/FPK) are the most widely used methods in the data analysis of RNA-seq as well as other sequencing data.

1.3 RESEARCH OBJECTIVES

As aforementioned, this thesis includes three major projects that are related to epigenetic, post-transcriptional, and translational regulation of gene expression. The first project is "DNA methylomes of bovine gametes and in vivo produced preimplantation embryos". In this project, we characterize genome-scale DNA methylation of bovine sperm and individual in vivo developed oocytes and preimplantation embryos by using RRBS data. The findings in this study provide insights into the complex epigenetic regulation of gene expression in early embryos. The second project is "PRAS: Predicting functional targets of RNA binding proteins based on CLIP-seq peaks". In this project, we propose the Protein-RNA Association Strength (PRAS), which integrates the intensities and positions of the binding peaks of RBPs for functional mRNA targets prediction. Leveraging the position information of the binding peaks, PRAS is a bridge linking peak-calling methods and the interpretation of RBPs' biological functions, which strengthens the analysis of CLIP-seq data. The third project is "c-MYC regulates mRNA translation efficiency and start-site selection in lymphoma". In this project, we show the oncogenic c-MYC (MYC) alter the efficiency and start-site usage of mRNA translation in lymphoma by analyzing the ribo-seq and rna-seq data. Our findings in this project provide new insight into the biological activity of MYC and its effect on mRNA translation both globally and locally.

DNA methylation regulates gene expression in the epigenetic process, RBPs control gene expression in the post-transcriptional regulation, and MYC alters gene expression in the protein level in lymphoma. The findings in all these three projects provide novel insights into the complex biological mechanisms as a step forward to decipher the gene expression control.

CHAPTER 2: DNA METHYLOMES OF BOVINE GAMETES AND IN VIVO PRODUCED PREIMPLANTATION EMBRYOS

2.1 INTRODUCTION

Cytosine methylation is an important epigenetic modification that is largely restricted to CG dinucleotides and serves to regulate gene transcription for differentiation, gene imprinting, and X-chromosome inactivation (Li et al. 1993; Bird 2002; Jaenisch and Bird 2003). The most dramatic genome-wide methylation changes occur in primordial germ cells and during preimplantation development (Smallwood et al. 2011; Seisenberger et al. 2012; Smith et al. 2012; Guo et al. 2014; Smith et al. 2014; Gkountela et al. 2015; Guo et al. 2015; Gao et al. 2017). In early embryos, this involves the ultimate removal of cytosine methylation acquired in the gametes prior to fertilization, a process extensively characterized at the single-base level in the mouse, and more recently in humans (Smallwood et al. 2011; Smith et al. 2012; Guo et al. 2014; Smith et al. 2014). Interestingly, the demethylation of the two parental genomes happens differently. Immediately after fertilization, the 5-methylcytosine (5mC) in the paternal pronucleus is actively converted to 5-hydroxymethylcytosine (5hmC) by the enzyme tet methylcytosine dioxygenase 3 (TET3) (Inoue and Zhang 2011; Iqbal et al. 2011; Wossidlo et al. 2011). In contrast, the 5mC in the maternal pronucleus is largely protected from the actions of TET3 by Stella/Pgc7/Dppa3, which interacts with H3K9me2 that is enriched in the maternal pronucleus (Nakamura et al. 2007; Nakamura et al. 2012; Bakhtari and Ross 2014). The differentially methylated pronuclei fuse and methylation (5mC and 5hmC) levels decrease in a replication dependent manner during cleavage division ultimately reaching a nadir. This is followed by large-scale de novo methylation that sets the stage for differentiation (Dean et al. 2001; Santos et al. 2002; Petrussa et al. 2016). The gradual demethylation during preimplantation development has been observed in humans, mice, and cattle, but the timing of the major wave of genome-wide de novo methylation differs (Dean et al. 2001; Smallwood et al. 2011; Smith et al. 2012; Guo et al. 2014; Smith et al. 2014).

To date, the global characterization of the DNA methylation dynamics in embryos of domestic species remains at the immunostaining level (Dean et al. 2001; Dobbs et al. 2013; Zhang et al. 2016), with

the exception of a few stages of in vitro produced bovine embryos that were analyzed using the EmbryoGENE DNA methylation array (Salilew-Wondim et al. 2015). These studies primarily revealed the methylation of highly condensed repetitive DNA or the methylation level of sequences represented on the array. Although numerous studies have been performed to evaluate the methylation level of selected genes and regulatory regions during bovine embryo development (Dobbs et al. 2013; O'Doherty et al. 2015; Mattern et al. 2016; Urrego et al. 2017), the complete characterization of DNA methylation at the single-base level has not been reported. This characterization is critical to understanding the epigenetic reprogramming and regulation that occurs during normal, bovine embryonic development in vivo, and to providing insight into the epigenetic alterations that occur during in vitro maturation of oocytes and culture of embryos after in vitro fertilization. Environmental perturbations experienced during in vitro production are expected to influence the epigenetic reprogramming during this critical period, often leading to nonrandom epigenetic errors (Li et al. 2005; Fernandez-Gonzalez et al. 2010) that are linked to imprinting diseases in humans (Sutcliffe et al. 2006) and large offspring syndrome in ruminants (Young et al. 1998; Chen et al. 2013).

Reduced representation bisulfite sequencing (RRBS) detects clustered CGs that are mainly located in CpG islands (CGI) and are important for gene expression regulation. Here, we performed RRBS of bovine sperm, in vivo developed oocytes, and embryos from the 2-cell to the blastocyst stage, obtaining a comprehensive single-base resolution map of DNA demethylation dynamics across early bovine preimplantation development. This data resource is valuable because it provides the "gold standard" reference for embryos produced by assisted reproductive technologies, as well as identifying potential regulatory mechanisms of DNA methylation in gametes and during embryo development. Such a rich dataset from an economically important agricultural species not only provides evolutionary insights into the epigenetics of early development, but also serves as a good model for understanding potential causes of human infertility and the epigenetic effects of the assisted reproductive technologies that are designed to treat it.

2.2 MATERIALS AND METHODS

2.2.1 RRBS library and pre-processing

RRBS libraries were prepared from three replicates of each gamete and embryo stage and multiplexed, and sequenced in Illumina Hiseq2500 with 125 bp pair-end reads. In total, we analyzed 30 samples from 10 stages of development and obtained 165 Gb sequencing data.

Multiplexed sequencing reads were first trimmed to remove low-quality bases and adaptor sequences using the Trim_galore tool. The clean reads were then aligned to the bisulfite-converted reference bovine genome UMD3.1.1 (bosTau8) using Bismark alignment tools (Krueger and Andrews 2011) (version 0.16.3) with default parameters. Additionally, because the cytosines in a non-CG (CHH and CHG) context in the lambda DNA genome are definitely unmethylated, the lambda DNA was rebuilt as an extra reference for alignment and the calculation of the bisulfite conversion rate of each sample. Once the alignment was completed, the sorted BAM files were generated by Picard toolkit and a pileup file of mapped data was created for DNA methylation calculations.

2.2.2 Determination of methylation levels of CpG and non-CpG sites

The methylation levels of each sampled cytosine were estimated as the number of reported C ("methylated" reads) divided by the sum of reported C ("methylated" reads) and T ("unmethylated" reads) at the same positions of the reference genome. Every CpG site with read depth >1 was summed and counted in the total CpG coverage of the sample, only the CpG sites with at least fivefold read coverage were used to quantify the DNA methylation level of each sample. We then performed the 100-bp tile-based DNA methylation calculation algorithm (Smith et al. 2012). First, we binned the reference genome into consecutive 100-bp tiles. Then, the number of reported C divided by the sum of reported C and T captured in the 100-bp tiles was regarded as the 100-bp-tile averaged DNA methylation level. The DNA methylation level for a given sample was the average of the 100-bp tiles, while the DNA methylation level of a stage was the arithmetic average value of all biological replicates of that stage. The CpG density for every CpG

site was calculated as the total number of all CpGs 50 bp up- and downstream of that CpG site. The CpG density for every 100-bp tile was calculated as the averaged CpG number in the tile. The tiles with methylation level over or equal to 20% or below 20% were defined as high/intermediate or low methylated tiles, respectively. For non-CpG methylation, the same calculation strategy was used.

2.2.3 Characterization of genomic features

The annotated retroelements, such as long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), and long terminal repeats (LTRs), and their subfamilies were downloaded from the RepeatMasker track of UCSC genome browser. Promoters were defined as regions of 1000 bp upstream of transcription start site (TSS) of each gene. Other regions, such as CGIs, exons, and introns, were downloaded from UCSC tables with UMD3.1.1 track. The intragenic regions were sequences from TSS to transcription end site (TES), while the intergenic regions were defined as the complement of intragenic regions in the bovine genome. For each annotated genomic region, the DNA methylation level was calculated from the average of all CpG sites in the region with more than fivefold coverage. Additionally, when quantifying the DNA methylation level of promoters, only those with at least five CpG sites were retained for further analysis.

2.2.4 Identification of dynamically methylated tiles and gamete-specific differentially methylated regions

We systematically compared the DNA methylation levels of overlapped 100-bp tiles in each of the compared groups or consecutive stages. For example, we regarded 100-bp tiles as gamete-specific differentially methylated regions (DMRs) if the methylation levels of one type of gametes (such as the sperm) were greater than 75%, while the other type (such as MII oocytes) were less than 25%, with a Benjamini-Hochberg false discovery rate (FDR) corrected $P \leq 0.05$ from a two-sided Student t test. Additionally, if a 100-bp tile had absolute methylation change >40% between the compared groups with an FDR-corrected Fisher's exact test of $P \leq 0.05$, it was then classified as a changing tile, while the remaining

tiles were considered stable. The gene ontology analysis was done for the genes with DMRs using DAVID online (https://david.ncifcrf.gov) (Huang da et al. 2009).

2.2.5 Correlation of gene expression and DNA methylation

Our previously published gene expression data from bovine in vivo MII oocytes and preimplantation embryos, GSE59186 (Jiang et al. 2014) , and the raw data of in vitro MII oocytes, GSE52415 (Graf et al. 2014), were pooled and aligned using Tophat2 (Kim et al. 2013) against bovine genome UMD3.1.1 (bosTau8) with default settings. Cufflinks (Trapnell et al. 2012) was used for quantification of FPKM values with default settings. The log2 of the gene expression levels (FPKM) of detectable genes (FPKM >0.1) and the DNA methylation levels of the promoters of each corresponding expressed gene were calculated. Finally, Spearman correlation coefficients (r) between gene expression and DNA methylation levels of promoters were calculated and plotted in R package.

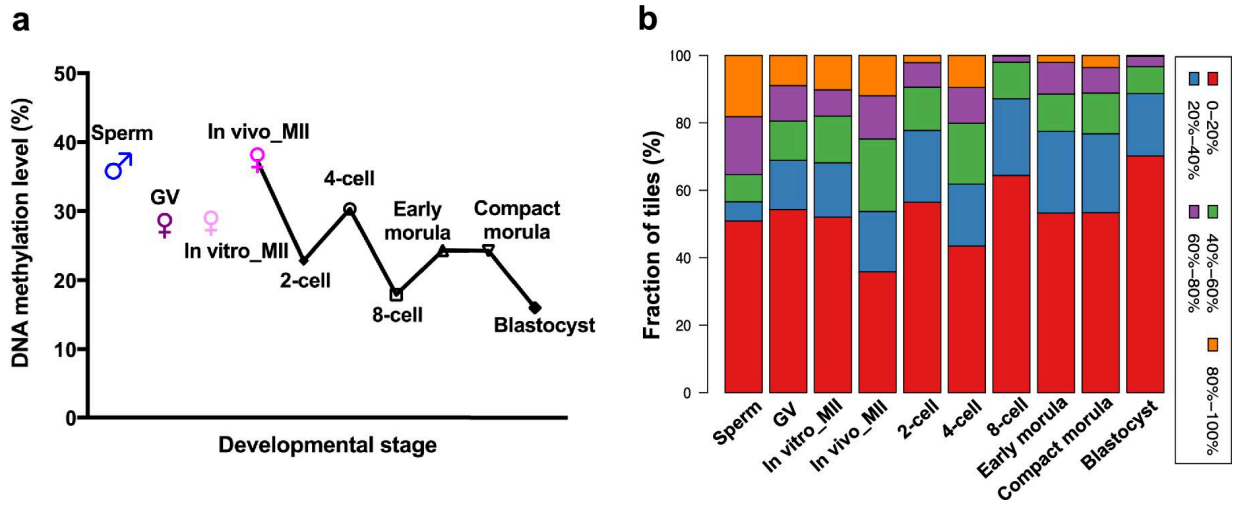


Figure 2.1 Bovine preimplantation embryos undergo genome-wide DNA demethylation. (a) The overall methylation levels of bovine gametes and early embryos. The averaged DNA methylation level was calculated based on the common 100-base-pair (bp) tiles detected in all stages analyzed. (b) Histogram of the fractions of tiles with 0–20%, 20–40%, 40–60%, 60–80%, and 80–100% methylation levels across different developmental stages.

2.3 RESULTS AND DISCUSSION

2.3.1 Genome-wide DNA demethylation in bovine preimplantation embryos

Using RRBS, we analyzed a total of 30 samples that included bovine sperm and individual in vivo matured oocytes and embryos of 10 different developmental stages (n = 3) and obtained 165 Gb of sequencing data (Supplementary Table S2.1). The raw FASTQ files and processed methylation calling files are available at Gene Expression Omnibus (GEO) (www.ncbi.nlm.nih.gov/geo) under the accession number GSE110400. After alignment, we removed three samples with extremely low mapping efficiency (Supplementary Table S2.1). The bisulfite conversion efficiency, estimated from spiked Lambda DNA, was more than 99% (Supplementary Table S2.1). RRBS provided the expected genomic coverage and reproducibility (Supplementary Table S2.1 and Supplementary Figure S2.1a). On average, we captured 2639,860 CpGs per stage for a total of 10X genome coverage (Supplementary Table S2.1). Captured CpGs were broadly spread across each chromosome (Supplementary Figure S2.1b). The overall DNA methylation level of each developmental stage was obtained by averaging methylated cytosines in 100-bp tiles of the reference genome. Using commonly detected tiles across all stages, the methylation could be divided into two distinct profiles (Supplementary Figure S2.1a and Figure 2.1a and b): (1) highly methylated sperm and in vivo matured oocytes and (2) cleavage-stage embryos with reduced methylation, the lowest level observed at the blastocyst stage (Figure 2.1a and b). These patterns are similar to those found in mice (Smallwood et al. 2011; Smith et al. 2012; Wang et al. 2014a) and humans (Guo et al. 2014; Smith et al. 2014). Two additional profiles are also noteworthy: from the 2-/4-cell to 8-cell stage, and from the early/compact morula to the blastocyst stage (Supplementary Figure S2.1a). The demethylation during early preimplantation development has been consistently reported in 5mC immunofluorescence-based studies (Dean et al. 2001; Fulka et al. 2004; Dobbs et al. 2013; Zhang et al. 2016). However, we did not observe the de novo methylation in blastocyst stage embryos that has been observed by immunofluorescence. This could be due to several factors, chief among them the analysis of in vivo embryos in this study. In vitro production of embryos has long been known to perturb epigenetic modifications (Reis e Silva et al. 2012; Canovas et al. 2017). Another potential cause of the discordance could be the difference in resolution

between immunofluorescence and RRBS, with the former providing a localized, low-resolution view of highly methylated regions in the embryo. Lastly, it could be due to technical attributes inherent to immunofluorescence (Salvaing et al. 2015). Indeed, de novo methylation at the blastocyst stage was also not observed by high-throughput sequencing analyses (RRBS or MethylC-seq) of mouse (Smallwood et al. 2011; Wang et al. 2014a) or human embryos (Guo et al. 2014; Okae et al. 2014; Smith et al. 2014).

Despite similarities to the patterns observed in mouse and human embryos, distinct features in the bovine embryos were observed. Specifically, a dramatic decrease in DNA methylation occurred between gametes and the 2-cell stage, with the average DNA methylation level decreasing from 37% in the sperm and in vivo MII oocytes to 22% in the 2-cell embryos (Figure 2.1a, Supplementary Figure S2.1c). Of interest, a significantly lower methylation level was found in the X chromosome compared to autosomes in sperm samples (Supplementary Figure S2.1c). A further and major reduction of DNA methylation to around 18% occurred as the embryo progressed to the 8-cell stage, coinciding with major embryonic genome activation (Misirlioglu et al. 2006; Kues et al. 2008; Graf et al. 2014; Jiang et al. 2014) (EGA; Figure 2.1a, Supplementary Figure S2.1c). This is consistent with the loss of methylation over multiple cleavage divisions due to the absence of the maintenance DNA methyltransferase 1 (DNMT1) (Kurihara et al. 2008), and to major increases in transcription from the embryonic genome (Braude et al. 1988; Hamatani et al. 2004; Wang et al. 2004; Yan et al. 2013; Graf et al. 2014; Jiang et al. 2014). Subsequently, an even more dramatic decrease was seen between the morula and blastocyst stages (16%; Figure 2.1a, Supplementary Figure S2.1c). The timing of this second demethylation wave correlated with the differentiation of the trophectoderm and inner cell mass, which involves the activation of specific genes, such as POU class 5 homeobox 1 (POU5F1) and caudal type homeobox 2 (CDX2) (Niwa et al. 2005; Strumpf et al. 2005; Sakurai et al. 2016). This is also correlated with the increased DNA methyltransferase 3 alpha (DNMT3A) expression in blastocysts (Jiang et al. 2014). In contrast, a moderate increase in DNA methylation was found from the 2- to 4-cell, and from 8-cell to early/compact morula, which is consistent with de novo methylation due to the elevated DNA methyltransferase 3 beta (DNMT3B) and DNMT3A expression in 4-cell and 16-cell stage bovine embryos, respectively (Jiang et al. 2014). Also of considerable interest, a

significant difference in the methylation level between in vivo and in vitro matured oocytes was found (Figure 2.1a and b). In vitro maturation maintained GV oocyte levels of methylation while in vivo maturation increased DNA methylation levels; this is consistent with the observation that in vitro maturation produces about 75% nuclear maturation, but cytoplasmic maturation is much more incomplete (Watson 2007). As a result, in vivo-derived oocytes have a higher developmental potential compared to in vitro (Rizos et al. 2002; Smith et al. 2009). These observations provide the underlying mechanism for the abnormal gene expression and reduced embryo and fetal development when oocytes are matured in vitro.

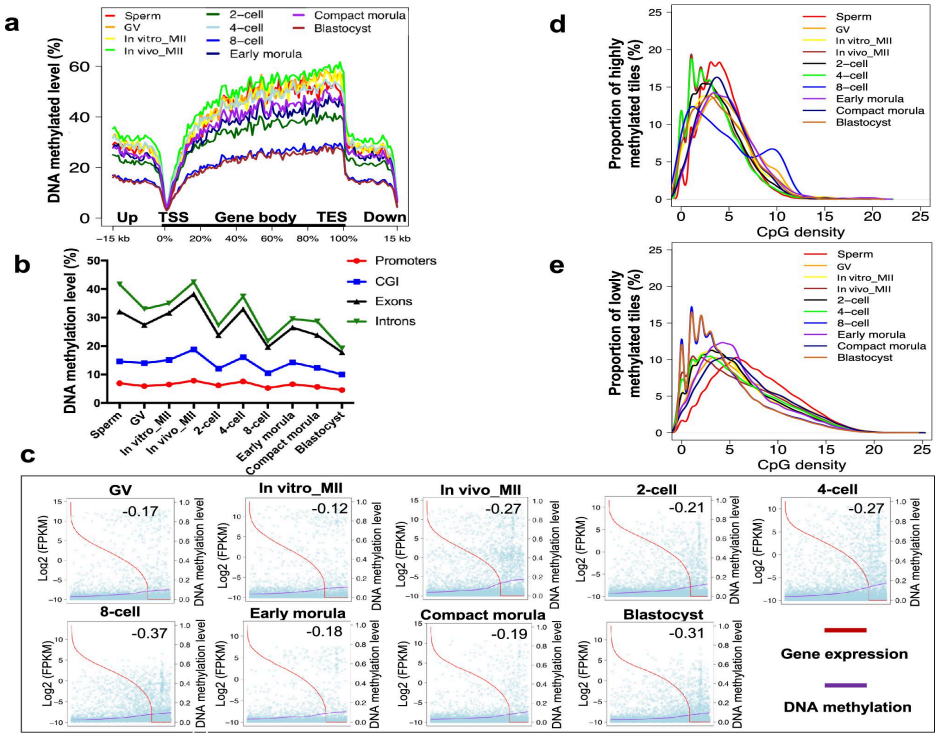


Figure 2.2 Characteristics of DNA methylation patterns during bovine early embryonic development. (a) Averaged DNA methylation levels along the gene bodies and 15 kilobases (kb) up- and downstream of the transcription start sites (TSS) and transcription end site (TES), respectively, of all reference genes. (b) The averaged DNA methylation level of each developmental stage on annotated bovine genome features, including promoters, CGI, exons, and introns. (c) Pearson correlation coefficients (r) of DNA methylation levels of promoter regions (purple curve) and the relative expression levels of the corresponding genes (red curve). The log2 of the gene expression levels from RNA-seq (reads per kilobase per million, RPKM) was calculated and presented. (d and e) The distribution of highly (>20%) and lowly (<20%) methylated tiles at each developmental stage against CpG density, respectively.

2.3.2 Genome-wide methylation maps and correlation with gene expression

As seen in mouse and human embryos (Smith et al. 2012; Guo et al. 2014), a distinct methylation pattern was observed in and around all annotated gene bodies, which were progressively more methylated than the 15-kb intergenic regions both up- and downstream. Specifically, the TSS (or 0%; Figure 2.2a) was associated with a sharp decline in methylation. The methylation then gradually increased in the gene body and plateaued until another sharp decline was observed at the TES (or 100%) (Figure 2.2a), which brought the DNA methylation close to the level of TSS. These patterns suggest that DNA methylation may be used by cells as a unique marker for gene boundaries. Overall, promoter and CGI regions were significantly lowly methylated compared to exons and introns (Figure 2.2b) across all stages. This may be necessary because early bovine embryos express on average 10,000 genes (Jiang et al. 2014), much higher than most differentiated tissues. Using our RNA sequencing (RNA-seq) data of early in vivo developed bovine embryos (Graf et al. 2014; Jiang et al. 2014), we found negative correlations between promoter methylation and the expression levels of corresponding genes during preimplantation development (Figure 2.2c), especially after EGA at the 8-cell stage and later in the blastocyst ($r<0.3$; Figure 2.2c, Supplementary Table S2.2). Because promoters, and to a large extent CGI, hardly changed their methylation levels across development, and all the methylation changes were in fact associated with exons and introns (Figure 2.2b), we also performed correlation analysis between CGI or exon methylation and the expression levels of corresponding genes (data not shown); similar negative correlation patterns were found with CGI or exons as in the promoter regions.

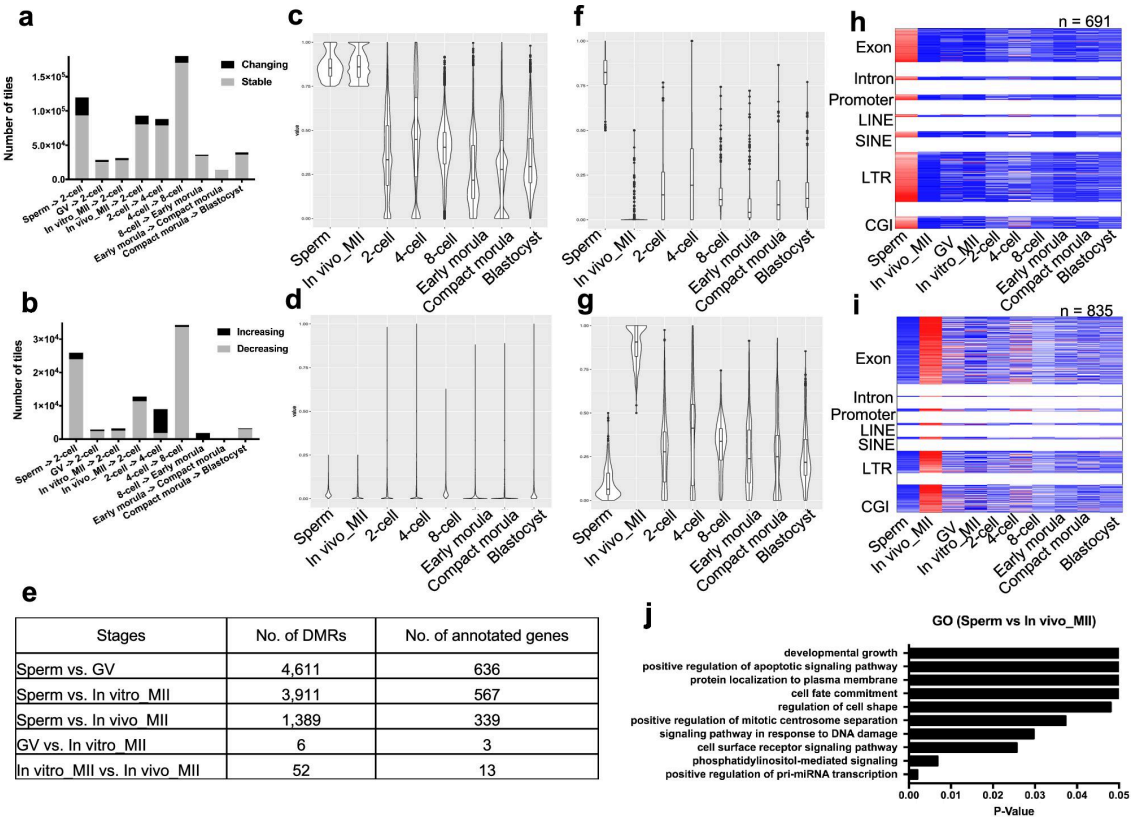
The bovine gametes and preimplantation embryos exhibited an inverse relationship between CpG density and methylation levels (Figure 2.2d and e); regions with high CpG density tended to be hypomethylated (<25%) and vice versa (>75%; Figure 2.2d and e), consistent with the patterns observed in mouse and human embryos (Smith et al. 2012; Guo et al. 2014). Interestingly, this correlation was more visible in gametes, and less so in the 8-cell and blastocyst stage embryos (Supplementary Figure S2.2), coinciding with EGA at the 8-cell stage and early lineage specification in blastocysts.

Although methylation mainly occurs at CpG sites, non-CpG methylation has been reported in oocytes (Tomizawa et al. 2011; Shirane et al. 2013; Guo et al. 2014) and human embryonic stem cells (Lister et al. 2009). We also observed reduced but detectable levels of non-CpG methylation in bovine early embryos (Supplementary Figure S2.3a). Because RRBS detects methylation mainly in CGIs, the non-CpG methylation identified was also located within CGIs. CpG and non-CpG methylation had similar enrichment patterns in and around gene bodies (Supplementary Figure S2.3b); this was also observed in human oocytes (Guo et al. 2014). Interestingly, extremely low non-CpG methylation was found in the sperm, 8-cell, and blastocyst stage embryos (Supplementary Figure S2.3a), all of which had no enrichment in gene bodies (Supplementary Figure S2.3b); the low non-CpG methylation was also found in human sperm and embryos (Guo et al. 2014). Additionally, there was no correlation between the levels of non-CpG methylation around gene bodies and expression in bovine gametes and early embryos ($r < 0.01$; Supplementary Figure S2.3c). It has been shown that non-CpG methylation regulates the expression of some genes (e.g. pyruvate dehydrogenase kinase 4 (PDK4)) (Barres et al. 2013) while not others (e.g. PPARG coactivator 1 alpha (PGC-1α)) (Barres et al. 2009). Since non-CpG methylation is prevalent only in specific tissues and cell types, or only in particular regions of the genome, its functional role remains unknown (Patil et al. 2014).

2.3.3 Signatures of differential methylation in bovine gametes

Although the overall levels of methylation underwent dramatic changes across embryonic stages (Figure 2.1a), the actual numbers of tiles with changed DNA methylation between consecutive stages were only minor compared to the number of stable ones. The greatest number of changed tiles was found between gametes and early cleavage embryos, and between the 4- and 8-cell stage embryos (Figure 2.3a) and corresponded to the overall level changes shown in Figure 2.1a. These changing tiles were likely involved in the activation of gene expression from the embryonic genome at these stages as we reported previously (Jiang et al. 2014). Notably, most regions in the genome showed reduced methylation in the first wave of EGA (between MII oocytes and the 2-cell stage) and major wave of EGA (between 4- and 8-cell embryos;

Figure 2.3b), likely from replication-dependent dilution and the lack of DNMT1 (Smith and Meissner 2013). From early morula to blastocyst stage, there was a minor increase in the number of tiles with reduced methylation (Figure 2.3a and b), consistent with the increase of methyltransferase expression (DNMT3A, B) during this transition (Okano et al. 1999; Smallwood et al. 2011; Smith and Meissner 2013; Jiang et al. 2014).



Stages	No. of DMRs	No. of annotated genes
Sperm vs. GV	4,611	636
Sperm vs. In vitro_MII	3,911	567
Sperm vs. In vivo_MII	1,389	339
GV vs. In vitro_MII	6	3
In vitro_MII vs. In vivo_MII	52	13

Figure 2.3 Major transitions in DNA methylation during bovine early development and key features of gamete-specific differentially methylated regions (DMRs). (a) The number of common tiles between gametes and consecutive stages that changed (black) or were stable (gray) in DNA methylation. (b) The number of common tiles between gametes and 2-cell embryos or embryos at consecutive stages of development that either had increased (black) or decreased (gray) methylation levels. DNA methylation levels in early embryos for tiles hypermethylated (c) and hypomethylated (d) in sperm and in vivo matured oocytes. (e) The number of DMRs and the number of corresponding genes between gametes of different types. DNA methylation levels of DMRs specific for sperm (f) and in vivo matured oocytes (g) in early embryos. Heatmaps of methylation levels (blue to red = low to high) for DMRs specific for sperm (h) and in vivo matured oocytes (i) in early embryos. Only DMRs that are significantly different ($P < 0.05$) between sperm and in vivo MII oocytes are presented. (j) Top represented gene ontology (GO) terms enriched in genes that had differential methylation between sperm and in vivo matured oocytes.

We next examined the similarities in DNA methylation between sperm and oocytes. Sperm and MII oocytes showed comparable methylation patterns for most covered tiles, which were either hypermethylated (methylation level ≥75%) or hypomethylated (≤25%) in both gametes (Figure 2.3c and d, Supplementary Figure S2.4a–d). The hypermethylated regions appeared gradually demethylated across cleavage stages, while the hypomethylated regions remained relatively unchanged (Figure 2.3c and d, Supplementary Figure S2.4a–d). We further found that regions that were commonly hypermethylated in both sperm and oocytes were enriched in LINEs and SINEs as well as introns (Supplementary Table S2.3), suggesting that hypermethylated regions in bovine gametes probably mainly serve to repress the activity of transposable elements, and play a role in regulating alternative splicing (Sela et al. 2010; Lev Maor et al. 2015). In contrast, commonly hypomethylated regions were enriched in promoters and CGIs (Supplementary Table S2.3), suggesting these regions are important for the dynamic gene activity in embryogenesis. A similar phenomenon was seen in human embryos (Guo et al. 2014).

We also examined significantly ($P < 0.05$) DMRs between sperm and oocytes. In total, we identified 1389 DMRs between sperm and oocytes matured in vivo, which corresponded to 339 genes (Figure 2.3 e). Of note, sperm-specific DMRs, which were strongly enriched in LTRs (Figure 2.3 h), rapidly lost methylation to the background level by 2-cell stage. Oocyte-specific DMRs, however, were often localized to exons and CGIs (Figure 2.3 i) and demethylated gradually across cleavage stages (Figure 2.3 f and g, Supplementary Figure S2.4g–j). Consistent with the overall averaged methylation levels (Figure 2.1a), both oocyte- and sperm-specific DMRs were more methylated than the same DMRs in cleavage and blastocyst stage embryos (Figure 2.3 h and i). Gene ontology analysis showed that genes associated with gamete-specific DMRs were clearly enriched for active cellular functions, such as regulation of transcription, signaling pathway, cell shape, cell fate, and developmental growth (Figure 2.3j).

Because more developmental failures occur in conceptuses derived from in vitro matured oocytes than those matured in vivo (Leibfried-Rutledge et al. 1987; Rizos et al. 2002), we further investigated the differential DNA methylation between these two types of oocytes. A total of 52 DMRs ($P < 0.05$; Figure 2.3 e, Supplementary Table S2.4), associated with 13 genes, were found (Figure 1.3 e, Supplementary Table

S2.4). Interestingly, many of them have not been characterized for their roles in maturation, making them good candidates for gene-specific epigenetic modification studies. During in vitro maturation only six tiles, corresponding to three genes (Figure 2.3 e), carbonic anhydrase 6 (CA6), caspase recruitment domain family member 11 (CARD11), and espin like (ESPNL), changed their methylation from the GV stage.

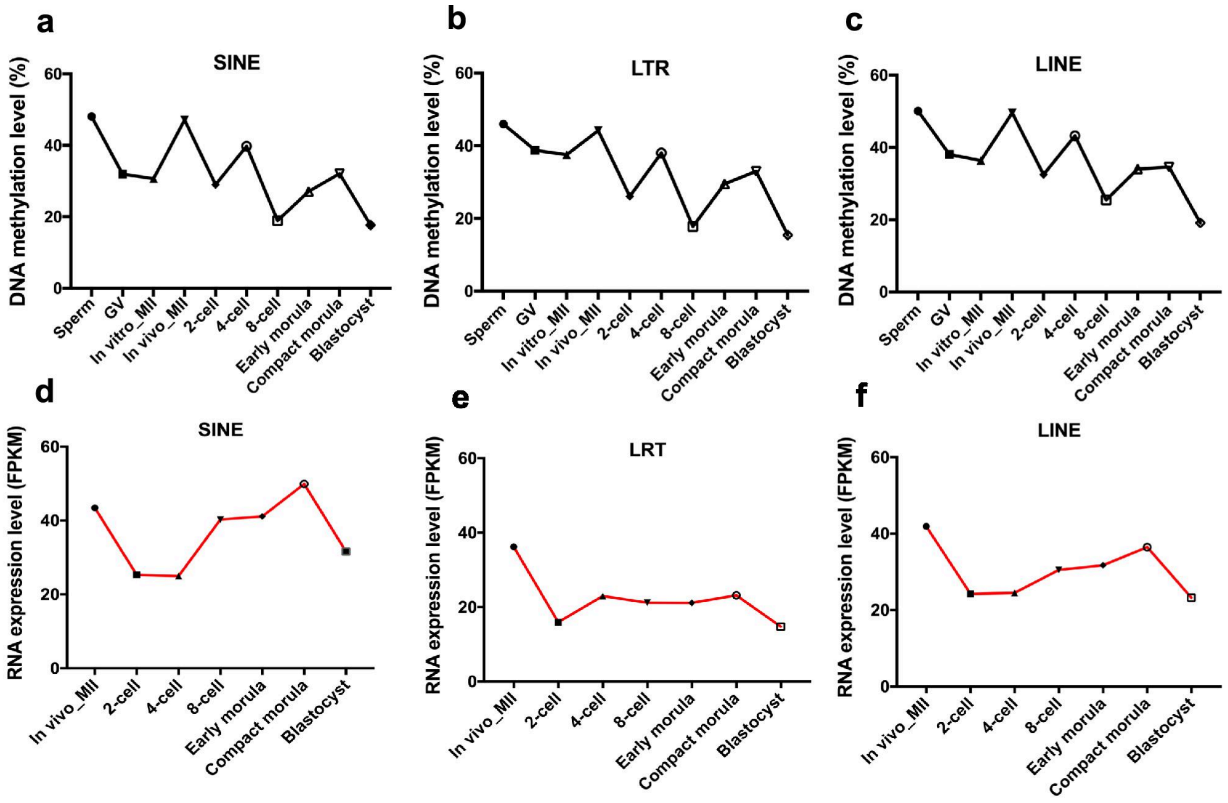


Figure 2.4 Dynamics of DNA methylation and expression patterns of transposable elements. DNA methylation levels of short interspersed nuclear elements (SINEs; a) long terminal repeats (LTRs; b), and long interspersed nuclear elements (LINEs; c). Relative expression levels of SINEs (d), LTRs (e), and LINEs (f).

2.3.4 Dynamics of DNA methylation and expression patterns of transposable elements

SINEs represent the majority of bovine genome repetitive content with LINEs being the second most prevalent (Adelson et al. 2009). Since we uncovered high levels of methylation of these sequences in the gametes, we were interested in determining the correlation of DNA methylation with the expression of SINEs and LINEs across development. Regardless of the sequences, the transposable elements had similar DNA methylation dynamics to what was seen for the overall genome methylation, i.e. higher methylation in sperm and in vivo matured oocytes, followed by demethylation at the 8-cell stage, reaching a nadir in blastocysts (Figure 2.4a–c). Transcription of all transposable elements, however, did not appear to follow the changes in DNA methylation except between the 2- and 8-cell stages (Figure 2.4d–f), where there seems to be a rough negative correlation of DNA methylation and transcription levels. It is possible that other mechanisms, or methylation not revealed by RRBS, are involved. Furthermore, evolutionary age of the transposable elements appeared to be correlated with methylation levels. For example, BovB (suggested to be old) had the highest methylation levels across all stages compared to other transposable elements (Supplementary Figure S2.5a). However, the observation that the evolutionarily younger L1 had a slightly higher methylation and transcription levels than L2 in bovine oocytes and early embryos (Supplementary Figure S2.5b) may suggest that young transposable elements are not demethylated to the same extent as their older counterparts.

2.3.5 DNA methylation dynamics of imprinted genes in bovine gametes and embryos

The mechanism of genetic imprinting often involves allele-specific DNA methylation in oocytes and sperm (Li and Sasaki 2011). To date, the DNA methylation profiles of bovine imprinted genes have not yet been well characterized in bovine gametes and across pre-implantation development, with the exception of small nuclear ribonucleoprotein polypeptide N (SNRPN), mesoderm specific transcript (MEST), PLAG1 like zinc finger 1 (PLAGL1), paternally expressed 10 (PEG10), insulin like growth factor 2 receptor (IGF2R), and insulin like growth factor 2 (IGF2) in sperm and oocytes (Gebert et al. 2006; O'Doherty et al. 2012), and SNRPN, MEST, PLAGL1, PEG10, IGF2, and imprinted maternally expressed

transcript (H19) in day 7 blastocysts (Gebert et al. 2009; O'Doherty et al. 2015). We assessed the methylation levels of all 29 genes known to be imprinted in the bovine (http://www.geneimprint.com/site/genes-by-species.Bos+taurus) (Tian 2014; Chen et al. 2015; Jiang et al. 2015). Only 15 were well covered by the captured CpGs using RRBS. Five maternally imprinted genes (neuronatin (NNAT), MEST, PLAGL1, PEG10, and SNRPN) had higher DNA methylation in in vivo matured oocytes than sperm (Figure 2.5). Conversely, three paternally imprinted genes (H19, maternally expressed 3 (MEG3), and tumor-suppressing subchromosomal transferable fragment 4 (TSSC4)) were more methylated in the sperm (Figure 2.5). As expected, the methylation levels for these eight imprinted genes in cleavage stage embryos were half of the high levels observed in gametes (Figure 2.5). These DMRs are good candidates for further study to determine if they are imprinting control elements. Our results not only confirmed that a number of bovine imprinted genes contain allele-specific methylated regions (Lucifero et al. 2006; O'Doherty et al. 2012), but also provide evidence that the methylation in these regions resisted the global demethylation process in early embryonic development as anticipated (Bartolomei 2009).

Interestingly, the methylation patterns of three maternally imprinted genes, retrotransposon Gag like 1 (RTL1), sarcoglycan epsilon (SGCE), and IGF2, and four paternally imprinted genes, diacylglycerol O-acyltransferase 1 (DGAT1), pleckstrin homology like domain family A member 2 (PHLDA2), IGF2R, and GNAS complex locus (GNAS), had the opposite methylation patterns than expected (Figure 2.5). It is worth noting that the levels of methylation for SGCE and PHLDA2 were low (Figure 2.5). A previous study also reported higher methylation of IGF2 (Gebert et al. 2006) in sperm than in oocytes. In addition, most imprinted genes are clustered and controlled by imprint control regions in mice and seven of the clusters have been well characterized, including IGF2, IGF2R, and GNAS (Wan and Bartolomei 2008; Barlow and Bartolomei 2014). Experiments have also indicated that the DMRs have different effects in these three clusters and suggested that knowing the position of the DMR with respect to the imprinted genes in each cluster is essential for understanding their exact regulation mechanisms (Barlow and Bartolomei 2014). Moreover, the mRNA expression of all these genes, except for GNAS, was extremely low in bovine

gametes and early embryos (Jiang et al. 2015); therefore, it is likely that the imprinting regulation of such genes involves other epigenetic mechanisms (Bartolomei and Tilghman 1997; Jiang et al. 2015).

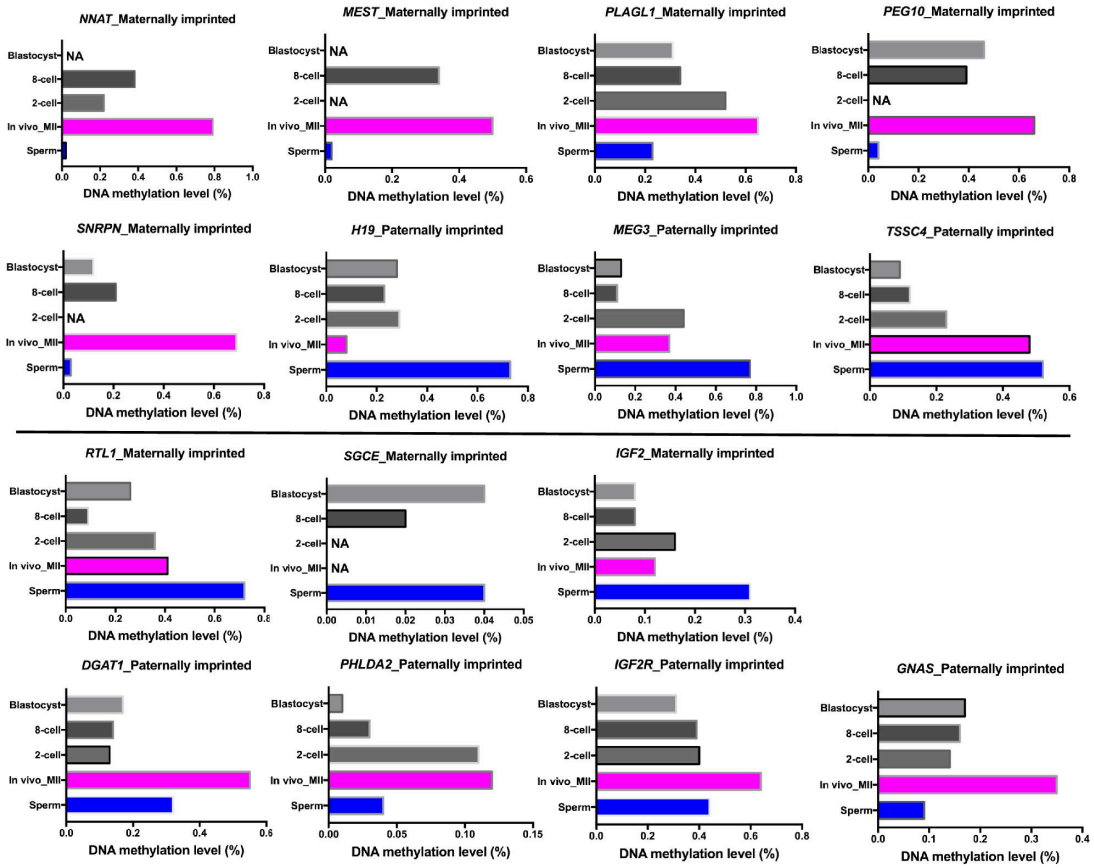


Figure 2.5 Methylation patterns of imprinted genes in bovine gametes. The DNA methylation levels of 15 known paternally and maternally imprinted genes in bovine gametes and early embryos. NA: methylation sites were not detected in the regions.

CHAPTER 3: PRAS: PREDICTING FUNCTIONAL TARGETS OF RNA BINDING PROTEINS BASED ON CLIP-SEQ PEAKS

3.1 INTRODUCTION

RNA-binding proteins (RBPs) are essential in many post-transcriptional regulatory processes, such as alternative splicing, stability, localization and editing (Keene 2007). For example, RBP Quaking plays important roles in pre-mRNA splicing and mRNA export (Chenard and Richard 2008); RBP HuR is an mRNA stability and splicing regulator (Lebedeva et al. 2011); RBP Ataxin-2 promotes mRNA stability and protein expression (Yokoshi et al. 2014). RBPs achieve their functions via binding to RNAs; therefore, it is of vital importance to study RNA-protein interaction. Cross-linking and immunoprecipitation followed by sequencing (CLIP-seq) approaches have been widely used to detect the binding peaks of RBPs at the transcriptome scale (Licatalosi et al. 2008; Hafner et al. 2010; Konig et al. 2010; Konig et al. 2012; Cook et al. 2015). Thus, the examination of CLIP-seq peaks informs us of the functional targets of RBPs.

Existing computational approaches for analyzing CLIP-seq data focus on detecting RBP binding peaks (Althammer et al. 2011; Corcoran et al. 2011; Uren et al. 2012; Lovci et al. 2013; Chen et al. 2014; Moore et al. 2014; Wang et al. 2014b; Wang et al. 2014c; Comoglio et al. 2015; Shah et al. 2017) or differential RBP binding peaks between two different conditions (Althammer et al. 2011; Uren et al. 2012; Erhard et al. 2013; Wang et al. 2014c). Computational methods for predicting the functional consequence of RBP binding peaks are less well-established (Mukherjee et al. 2011; Modic et al. 2013; Rot et al. 2017). Some studies suggest that the binding preferences of RBPs are associated with their specific functions. For example, HuR binding preferentially occurs close to the 3' splicing site, which is consistent with its known function on alternative splicing (Lebedeva et al. 2011); Ataxin-2, an mRNA stability regulator, has a tendency to bind close to the polyadenylation site (Yokoshi et al. 2014). A recent study revealed that RBP TDP-43 regulates poly(A) site usage in a position-dependent way (Rot et al. 2017).

In this paper, we develop a new approach named Protein-RNA Association Strength (PRAS), which incorporates the intensity and positional information of CLIP-seq peaks to quantitate the association between an RBP and its targets. We apply PRAS to study two CUGBP ELAV-like family proteins, CELF4

and CELF1 with both CLIP and perturbation RNA-seq data available. CELF4 (also known as Brunol4) is expressed as an mRNA regulator in the central nervous system across species (Meins et al. 2002; Yang et al. 2007). The deficiency of CELF4 is associated with a complex neurobehavioral disorder including seizures and autism-like features in human (Halgren et al. 2012; Barone et al. 2017) and in mice (Wagnon et al. 2011). iCLIP studies revealed that CELF4 preferentially binds, almost exclusively in 3' untranslated regions (UTRs), to mRNAs encoding many important neurological functions, (Wagnon et al. 2012). CELF1 is implicated in myotonic dystrophy (Timchenko et al. 1996). CELF1 is highly expressed in early embryonic stages and are then down-regulated dramatically in skeletal muscle and the heart during development (Ladd et al. 2005; Kalsotra et al. 2008). CELF1 has been reported to promote transcript deadenylation and the abnormal up-regulation of its protein level could contribute to the myotonic dystrophy pathology (Moraes et al. 2006; Wang et al. 2015). A more refined understanding of the functional targets of CELF RBPs is essential for understanding the impact of CELF in development and diseases, and may provide clues as to the mechanisms by which CELF impacts mRNA function. In addition, to demonstrate the robustness of PRAS, we examined its performance of detecting the functional targets in a large-scale collection of eCLIP data of RBPs in the integrated encyclopedia of DNA elements in the human genome (ENCODE). By applying PRAS to the eCLIP peaks of the RNA decay regulators, we demonstrate that PRAS outperforms other existing methods and also provide deeper understanding in the post-transcriptional regulation of these RBPs.

3.2 DESIGN AND IMPLEMENTATION

3.2.1 The framework of PRAS

The basis of PRAS is to score a potential functional target of an RBP based on both the intensities and positions of its binding sites. Our pipeline of calculating PRAS is shown in Figure 3.1. First, given a CLIP-seq dataset, the significant cross-linking sites that are within a small interval of each other (default: 20 nt) are merged as RBP binding peaks. If the called binding peaks are provided, we will use them directly. Second, if a reference position is provided by the user based on known knowledge of the function of the

RBP, PRAS will use it directly; if no reference position is given, PRAS will set it based on the RBP's binding preference, e.g., the distal end of the 3' UTR of the transcript (aka polyadenylation site). Finally, each transcript is scored as the sum of the intensities of the binding peaks weighted by the distances between the mid points of the binding peaks and the preselected reference position. All mRNAs are then ranked by the PRAS scores and can be tested for associations with functions.

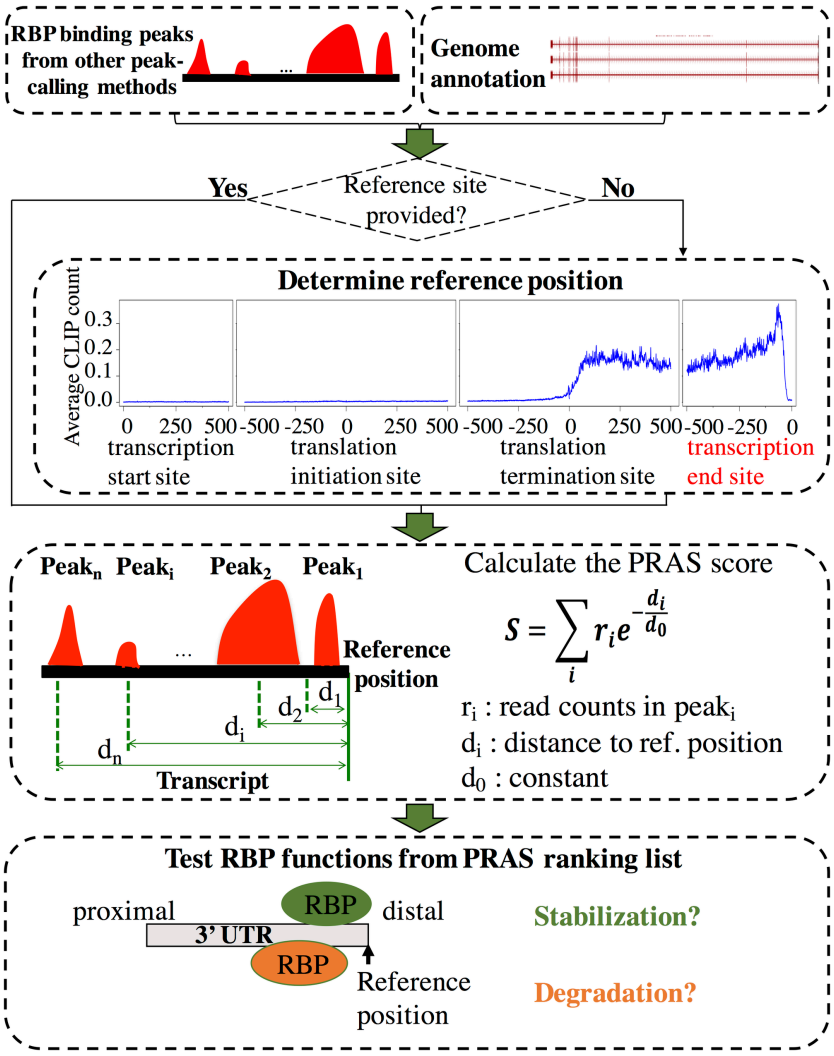


Figure 3.1 Flowchart of the PRAS pipeline. There are mainly three steps in calculating the PRAS scores. First, we merge the significant cross-linking sites as the binding peaks. Then, we use user-provided or automatically selected reference position and score each transcript based on both the intensities and the positions of the binding peaks. Finally, we rank the targets by PRAS and test RBP functions by independent datasets. The details of the PRAS calculation are described in the following section.

3.2.2 PRAS score calculation

As described in Fig 1, the PRAS score is based on the weighted sum of the intensities of the binding given detected CLIP-seq peaks. In the study that analyzed the interaction between DNA and proteins with ChIP-seq datasets, the exponential decay function was used to characterize the decreasing effects of a transcription factor binding peak on its targets with increasing distances (Ouyang et al. 2009). Therefore, we here construct the score to describe the regulatory effect of an RBP on its targets in a similar way. Specifically, we define the PRAS score for an mRNA as:

$$S = \sum_i r_i e^{-d_i/d_0}, \quad (1)$$

where r_i is the intensity (CLIP-seq read counts) of the *i*th peak cluster of the RBP, d_i is the distance (number of nucleotides) between the reference position and the *i*th peak cluster, and d_0 is a constant. For both CELF4 and CELF1 in mouse, we set the reference position as the distal 3' UTR and the constant $d_0 = 1000$ nt. Note that $d_0 = 1000$ nt is the default setting, but not a hard-set option in PRAS. For the RNA decay regulators in human, we set the constant $d_0 = 500$ nt. The details of d_0 estimation for RBPs in mouse and human are described in the Results and Discussions sections.

3.2.3 PRAS implementation

PRAS is implemented in Python (version 2.7.14 or above) and R (version 3.3.2 or above) scripts and has minimum requirements for the inputs. To reformat the annotation file, PRAS takes use of gtfToGenePred, a toolkit from the UCSC Genome Browser (Kent et al. 2002). PRAS also uses BEDTools (Quinlan and Hall 2010) to efficiently obtain the overlapping between the binding sites and the annotation regions. The annotation file should be the Gene Transfer Format (GTF) format and the peak file (no special requirement for the peak caller) should be the Browser Extensible Data (BED) format as the required input files, which are both the standard file formats. Details of usage can be found on the instruction page of our website: https://github.com/ouyang-lab/PRAS.

3.3 RESULTS AND DISCUSSIONS

3.3.1 PRAS score is a strong predictor of PCR-validated mRNA targets of CELF4

CELF4 is expressed in excitatory neurons of the adult mouse brain, from which iCLIP data are available (Yang et al. 2007; Wagnon et al. 2011; Wagnon et al. 2012). We collected the significant cross-linking sites detected by iCount (http://icount.fri.uni-lj.si) with false discovery rate (FDR) less than or equal to 0.05. We conducted a metagene analysis involving all 9,193 mRNAs that are bound by CELF4 and noted an enrichment of iCLIP reads at the distal (3' end) versus proximal (5' end) 3' UTR (Supplementary Figure S2.1). This preference suggests a potentially functional role of CELF4 binding close to the polyadenylation site.

We calculated the PRAS scores for CELF4 binding mRNAs with the polyadenylation site as the reference position, which gives the binding sites closer to the polyadenylation site higher weights. We estimated the decay parameter d_0 in Equation (1) based on the strength of the peak intensity decay shown in Supplementary Figure 3.1. In detail, we defined the weighting formula as $w = e^{-d/d_0}$ according to Equation (1). The highest average peak density, 0.843, appears at 63 nt to the 3' end of 3' UTR and the average peak density at 1000 nt upstream to the 3' end of the 3' UTR is 0.285 (Supplementary Figure S3.1). We calculated w as the ratio between the average peak intensity at the 1000 nt upstream to the 3' UTR and that of the 3' end of the 3' UTR, which is 0.285/0.843 = 0.339. By plugging d = 937 nt (which is 1000 nt – 63 nt) and w = 0.339 into the weighting formula, we obtained the estimation of 866 nt for d_0, which is approximately the default of 1000 nt. For comparison, we applied the expressRNA procedure of Rot et al.

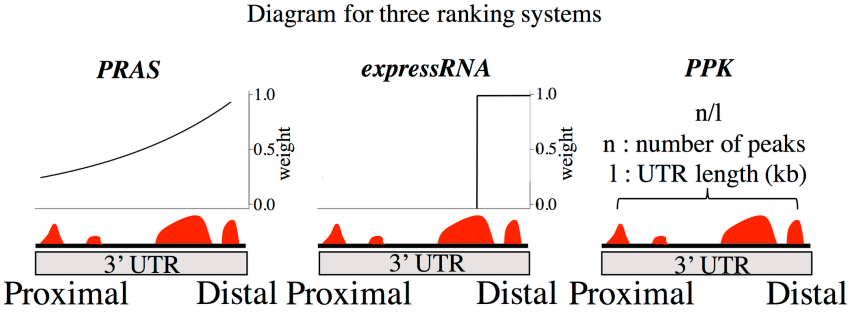


Figure 3.2 Diagram of three ranking methods.

(Rot et al. 2017), which sums the number of reads in CLIP peaks within 200 nt upstream and downstream flanking the polyadenylation sites (Figure 3.2). We also applied the procedure in Wang et al (Wang et al. 2015), which calculated the score as the number of significant CLIP peaks per kilobase (noted as PPK; Figure 3.2). Each of the three measurements ranks CELF4 binding mRNAs from high to low scores.

We then evaluated the performance of PRAS, expressRNA, and PPK on a list of known functional targets previously validated by qPCR in wild-type and *Celf4* null mouse brain, totaling 23 mRNAs (Wagnon et al. 2012). To investigate the ability of the three measurements to identify CELF4 functional targets, we performed receiver operating characteristic (ROC) analysis. We extracted the log fold change (LFC) of the qPCR values in *Celf4* null mouse brain over wild-type. The mRNAs with positive and negative LFCs were labelled as CELF4-degraded and CELF4-stabilized genes, respectively. The area under the curve (AUC) of the ROC curve was used to measure the prediction performance of the methods. We found that PRAS perfectly distinguished the PCR-validated CELF4-degraded and CELF4-stabilized genes (AUC=1), outperforming expressRNA (AUC=0.867) and PPK (AUC=0.7) (Figure 3.3a). This result suggests that given CLIP peaks, PRAS has greater ability to capture the functional targets of CELF4 compared to expressRNA and PPK. In addition, we examined the quantitative relationship between the PRAS scores and the qPCR LFCs of these known targets. A negative Pearson's correlation coefficient (-0.60) was obtained, suggesting that the more negative qPCR LFC a target has, the larger the PRAS score is (Figure 3.3b). The advantage of PRAS over expressRNA and PPK can be attributed to two factors. First, PRAS utilizes the binding bias of CELF4 towards the distal 3' UTRs of its validated targets (Figure 3.3c). expressRNA partially utilizes this bias by considering the 200 nt flanking region around the polyadenylation site, whereas PPK does not consider the binding bias. Second, unlike expressRNA which only considers a fixed flanking region, PRAS considers all binding peaks, which decreases loss of important RBP binding sites. The analysis of the validated targets of CELF4 suggests the importance of binding near the polyadenylation sites as a potential factor on how it regulates gene expression. By applying different decay parameter d_0 to PRAS, we found that PRAS obtained equally good performance over a reasonable range of d_0s (Supplementary Figure S3.2A and B). A d_0 that falls out of certain range will decrease the

performance of PRAS (Supplementary Figure S3A and B), because a too small d_0 can filter out the majority of iCLIP signals and a too large d_0 approximates the uniform weighting. The stable performance of PRAS with d_0 chosen around 1000nt shows the robustness of PRAS (Supplementary Figure S3).

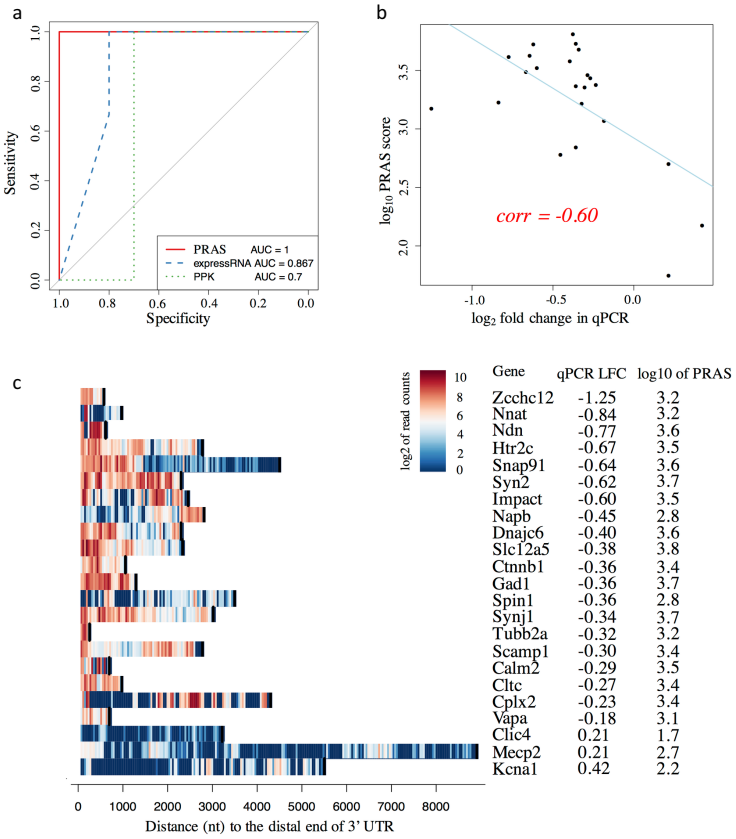


Figure 3.3. The qPCR-validated targets of CELF4. (a) The plot of ROC curves for PRAS, expressRNA, and PPK in the qPCR validated targets. The ROC analysis was done on the three methods' scores and the expression change. The corresponding ROC curves for PRAS, expressRNA, and PPK are indicated by red solid, blue dashed, and green dotted lines, respectively. The AUC of the corresponding ROC curves are listed at the bottom of the plot. (b) The scatter plot of the PRAS score against the qPCR log fold change (LFC). The X-axis represents the log2 fold change in qPCR from the wild-type to Celf4 null mouse brain. The Y-axis shows the log10 of PRAS score. Each black dot represents a validated target by the qPCR. The regression line is highlighted in blue color. The Pearson's correlation coefficient is indicated by the red text on the plot. (C) The heatmap of binding signals of the qPCR-validated targets. The X-axis represents the distance to the 3'end of the 3' UTR, and the Y-axis shows the genes in the validated list. The color shows the log2 of read counts of CELF4 iCLIP-seq within its significant peaks, where the warmer the color is the stronger the binding is. The black bars in each row shows the distance from the 5' end of the 3' UTR to the 3' end of the 3' UTR, which indicates the length of each 3' UTR.

3.3.2 PRAS score correlates with global mRNA change induced by CELF RBPs

To assess the ability of PRAS to detect RBP functional targets in the entire transcriptome, we extracted the top 500 genes ranked by permutation test *p*-values in the differential expression test between the wild-type and *Celf4* null mouse brain based on existing microarray datasets (Wagnon et al. 2012). We calculated the LFC for gene expression in *Celf4* null over wild-type mouse brain. The mRNAs have lower abundance (LFC < 0) in *Celf4* null genotype are more likely to be CELF4-stabilized targets, while the mRNAs with higher abundance (LFC > 0) in *Celf4* null brain were more likely to be CELF4-degraded targets. We sought to assess the ability of PRAS on capturing CELF4-stabilized vs. CELF4-degraded targets. Specifically, we first set a sequence of cutoffs as the quantiles (from 0.05 to 0.95 with step size as 0.05) of the distribution of the absolute value of the expression LFCs. Second, for each cutoff, we extracted a subset of genes whose absolute expression LFC is larger or equal to the cutoff. Finally, for each subset of potential CELF4 targets, we calculated the Spearman's correlation coefficient between the expression LFCs and the PRAS scores, in which the magnitude and sign of the correlation reflect the association between the two. For comparison, we also applied the same correlation analysis to expressRNA and PPK ranking scores. Line-charts of the Spearman's correlation coefficient of the three methods are shown in Figure 2.4a. We observed that the more stringent the expression LFC cutoff for the gene subset was set, the stronger the negative correlation between the PRAS score and the expression LFC was obtained, which suggests that PRAS is more powerful in capturing more reliable CELF4-stabilized targets. In addition, the expressRNA score is less correlated with the expression LFC, and the direction of the correlation between the PPK score and the expression LFC flips at different cutoffs. The results suggest that PRAS has greater ability to select the regulated mRNA targets compared to expressRNA and PPK.

We also extracted the top 500 genes ranked by their adjusted *p*-values in the differential expression test between the wild-type and *Celf1* over-expression in mouse muscle based on published RNA-seq datasets (Wang et al. 2015). In this dataset, mRNAs that have higher abundance upon *Celf1* over-expression (LFC > 0) are more likely to be CELF1-stabilized targets while those that have lower abundance upon *Celf1* over-expression (LFC < 0) were more likely to be CELF1-degraded targets. We evaluated the performance

of the three aforementioned methods using the same analysis as with CELF4. We used the 3' end of the 3' UTR as the reference site in PRAS to rank the mRNA targets based on the reported binding preference of CELF1 (Wang et al. 2015). PRAS has a stronger negative correlation with the expression LFC compared to expressRNA and PPK for each subset of the potential CELF1 targets (Figure 3.4b). These results suggest that PRAS is more powerful in capturing the reliable CELF1-degraded targets, consistent with the main regulatory function of CELF1 (Wang et al. 2015).

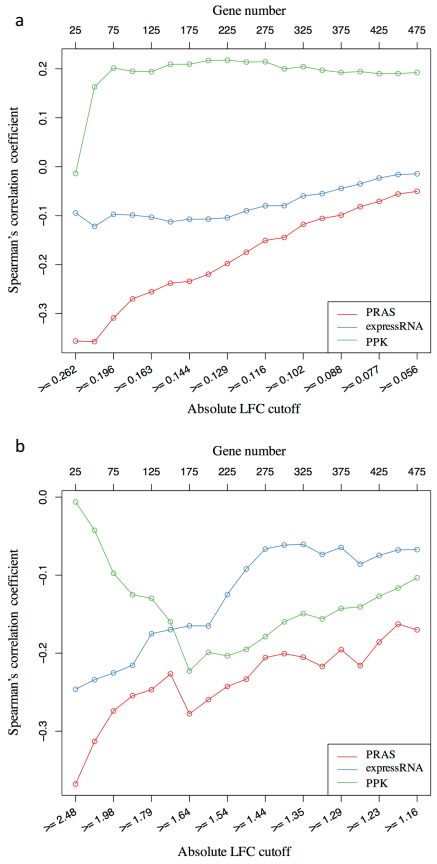


Figure 3.4 Correlation analysis between PRAS score and gene expression change. (a) The line-chart of Spearman's correlation coefficient between the gene score and the gene expression LFC in the Celf4-regulated list. The lower X-axis represents the different cutoffs applied to extract the subset of genes, the upper X-axis represents the number of genes corresponding to the applied cutoffs, and the Y-axis shows the value of Spearman's correlation coefficient. The corresponding curves for PRAS, expressRNA, and PPK are indicated by red, blue, and green lines, respectively. Each dot in the plot is for one subset of genes selected based on the absolute LFC cutoff. (b) Similar line-chart to A, but for the Celf1-regulated list. These two line-charts show that the higher ranked targets by PRAS have higher enrichment in the regulated lists comparing to the top ranked lists of expressRNA and PPK.

Next, we used different reference sites in PRAS for scoring functional targets of CELF4 and CELF1 in order to examine the effect of the reference site selection. We scored the targets of CELF4 using the 5' end of the 3'UTR as the reference site in PRAS (PRAS 5') and did a similar correlation analysis as above. We observed that the PRAS 5' score is also negatively correlated with the expression LFC and the magnitude of correlation improves with increasingly stringent cutoffs (Supplementary Figure S3.3A). However, the magnitude of the correlation is not as high as that of PRAS with the 3' end of 3' UTR as the reference site (PRAS 3') (Figure 3.4a). We also similarly analyzed the targets of CELF1 using PRAS 5'. Again, the PRAS 3' has stronger negative correlation with the expression LFC than PRAS 5' for the more reliable CELF1 targets (Supplementary Figure S3.3B). The results indicate that known biological knowledge can aid in reference site selection in PRAS for identifying the functional targets of the CELF proteins. The results also suggest that both the CELF4 and CELF1 proteins may regulate mRNAs via the distal 3' UTRs while having opposite effects on their targets. Indeed, this is plausible because CELF proteins play various roles in both co-transcriptional and post-transcriptional RNA regulation, as well as translation inhibition in different cellular contexts (Mukhopadhyay et al. 2003; Subramaniam et al. 2008; Dasgupta and Ladd 2012).

To examine the difference of taking the raw or the normalized read density of the CLIP peaks as the input of PRAS, we then used the Celf4 null iCLIP-seq as the negative control for the wild-type CELF4 iCLIP to score the functional targets of CELF4 with the 3' end of the 3'UTR as the reference site. Specifically, we replaced the iCLIP-seq read counts r_i in Equation (1) by the enrichment ratio $r_i \times log_2(\frac{r_i}{c_i})$ as suggested by Van Nostrand *et al* (Van Nostrand et al. 2018). We noted the PRAS score using the raw read intensity and the enrichment ratio of peaks as PRAS-raw and PRAS-norm, respectively. By applying the correlation analysis as above, we found that PRAS-norm has achieved stronger negative correlation with the expression LFC than PRAS-raw (Supplementary Figure S3.4). This improvement of performance indicates the important role of the negative control in reducing the noise, which is consistent with the results in (Van Nostrand et al. 2016). Even though PRAS-raw cannot achieve as good performance as PRAS-norm,

the difference in the performance between them is small (Supplementary Figure S3.4), which indicates that PRAS can handle the situation where the negative control of CLIP-seq is not available, such as the CELF1 data in our study.

3.3.3 PRAS identified targets are strongly enriched in functional categories

To further compare the functional relevance of the targets identified by PRAS, expressRNA and PPK, we performed gene ontology (GO) analysis on the top 500 mRNA targets of CELF4 ranked by each score (Figure 3.5a-c), which is similar to the analysis shown in Wagnon et al (Wagnon et al. 2012). There is much greater enrichment (5 to 40 orders based on p-values) of the categories related to suspected CELF4 function in the targets identified by PRAS than those identified by expressRNA and PPK. For example, in the class of "Biological Process", most of the top 10 significant categories for PRAS top-ranked targets are related to neuron or synaptic functions and ion transport, consistent with prior studies on CELF4 (Wagnon et al. 2012). These results suggest that PRAS captures CELF4 functional targets more precisely than the other methods being compared.

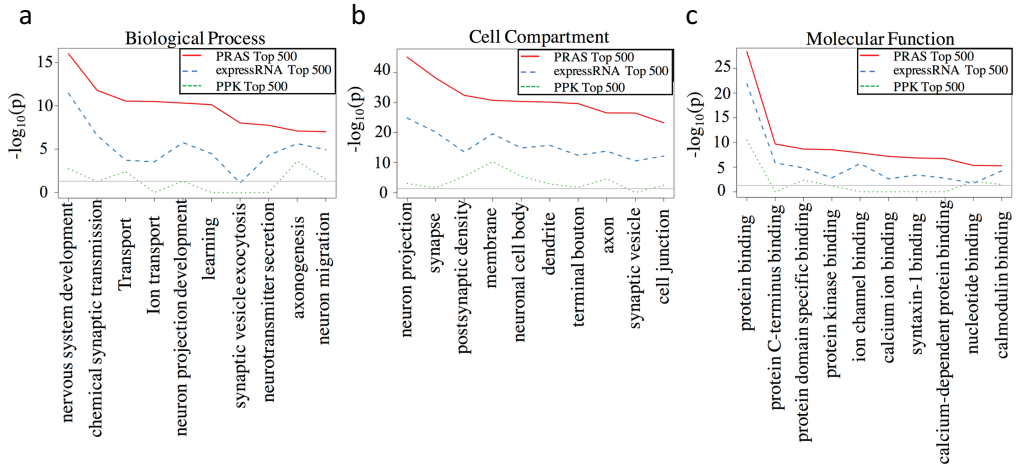


Figure 3.5 GO analysis of the top ranked targets in different methods and top differentially expressed genes. (a) "Biological Process" GO analysis line-chart. X-axis represents the GO term and Y-axis is the -log10(p-value) from the David GO analysis tool (https://david.ncifcrf.gov/) for the top 500 targets ranked by each method. PRAS is highlighted by red solid line, expressRNA is highlighted by blue dashed line, and PPK is highlighted by green dotted line. (b) Similar plot to A but for "Cell Compartment" GO analysis. (c) Similar plot to A but for "Molecular Function" GO analysis.

3.3.4 Functional targets are identified in a large-scale PRAS application to human RBPs

To demonstrate that PRAS has the potential for wide adoption, we further applied PRAS to the eCLIP data (Van Nostrand et al. 2016) in two human cell lines, K562 and HepG2, from the ENCODE consortium (Consortium 2012). Specifically, we selected the RBPs that are related to the RNA decay function (Van Nostrand et al. 2018) because this function can be clearly quantified at gene level in the differential expression (DE) analysis between the RBP knockdown and the wild-type RNA-seq samples. We collected the DE analysis results by DESeq (Love et al. 2014) from ENCODE and obtained 37 distinct RBPs, which include 28 and 32 RBPs in HepG2 and K562 cell line, respectively. We then applied PRAS to the eCLIP data using the enrichment ratio over the control sample described above as the peak intensities. In the parameter settings in PRAS, we selected the reference site for each RBP from 4 candidates: transcription start site, translation initiation site, translation termination site, and transcription end site, based on eCLIP peak intensity distribution along the transcript. Supplementary Figure S3.5 presents four example RBPs assigned with 4 different reference sites. To simplify the analysis, we applied $d_0 = 500\ nt$ to all the selected RBPs according to the distribution (Supplementary Figure S3.6) of the estimated decay parameters as described previously. This general selection of d_0 may not achieve the best performance of PRAS but is likely to be comparable with the best d_0 selection as discussed in the CELF4 data. After obtaining the PRAS scores, we did the correlation analysis of the DE (adjusted p-value <= 0.05) genes for each RBP. We found PRAS scores achieved significantly stronger correlation with the LFC in gene expression in comparison to expressRNA and PPK, with p-value equal to 3.8e-9 and 4.4e-4, respectively (Figure 3.6a). We then separated the RBPs by their reference site usage and found that the translation termination site and the transcription end site, both of which are related to the 3' UTR, constitute the majority of the RNA decay regulators' reference sites (Figure 3.6b). It suggests the essential association between the 3' UTR of transcripts and the regulation of their fates by RBPs. In addition, we found that the correlation can reflect important biological functions of RBPs. For example, the 5' poly(A) site (transcription end site) is used as the reference site for DDX6 in the HepG2 cell line (Supplementary Figure S3.5 C) and the PRAS score is negatively correlated with the LFC of DDX6's target gene expression (Figure

3.6b), which indicates that DDX6 may stabilize its targets via binding near to the poly(A) site. Interestingly, DDX6 is known to be an important regulator in mRNA decapping and degradation (Fenger-Gron et al. 2005; Hu et al. 2015), which supports our claim that PRAS has the ability to identify the biologically functional targets of the RBP regulators. All these results demonstrate that PRAS has the potential for wide adoption in RBP functional targets identification.

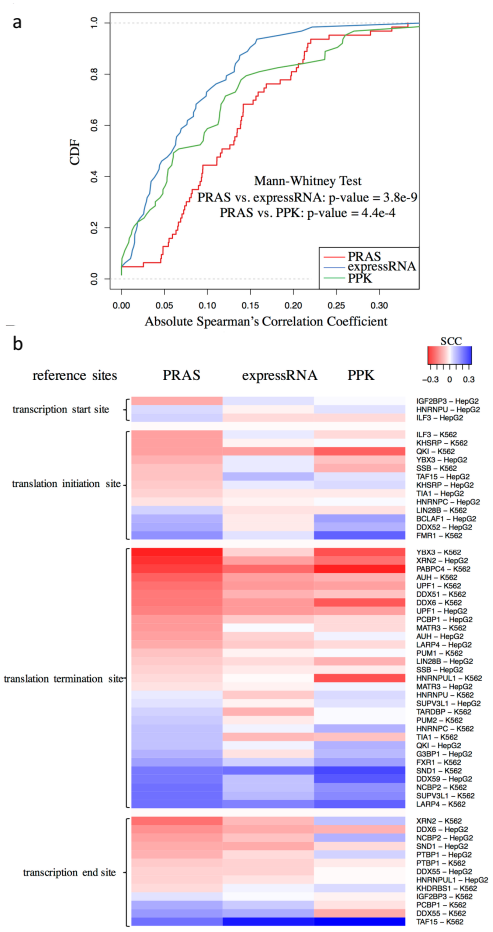


Figure 3.6 PRAS applied to RNA decay related RBPs. (a) The CDF curve of the absolute correlation coefficient between the gene score and LFC in gene expression. X-axis represents the absolute value of the Spearman's correlation coefficient between the gene score and LFC in gene expression (KO over wild-type). PRAS, expressRNA, and PPK is highlighted by red, blue and green line, respectively. The p-value of one-sided Mann-Whitney test is listed on the figure. (b) Heatmap of the Spearman's correlation coefficient. The Spearman's correlation coefficient between the gene score and LFC in gene expression for PRAS, expressRNA and PPK are listed from the left to the right. The values of the correlation coefficient are indicated by the color, where red and blue color indicates the positive correlation and the negative correlation, respectively. RBPs are grouped by their reference site usage and their ID and cell lines are listed at the right side.

3.3.5 Discussions on biological insights from the use of PRAS

In this study, we developed PRAS, a position dependent scoring method for identifying and prioritizing RBP functional targets. Weighting the proximity of RBP binding sites to a given reference position exponentially and combining the strengths of the binding signals, we obtained the PRAS scores and the ranking of all the mRNAs that have reliable binding sites of the RBP. We applied this approach to the iCLIP dataset of a neuronal disease-related RBP, CELF4 and to the CLIP dataset of a DM disease-related RBP, CELF1 – both belonging to the CELF family of RBP. We report a much stronger association between CELF4 and its targets at the distal 3' UTRs compared to internal 3' UTR positions. We also demonstrate that PRAS performs much better in predicting the mRNA targets stabilized by CELF4, compared to the other existing methods such as expressRNA and PPK. We further observe that PRAS performs much better at predicting the mRNA targets degraded by CELF1. These results not only suggest the importance of incorporating the positional information of the binding sites into target identification, but also suggest the important roles of the distal 3' UTRs in CELF protein regulated mRNAs.

The binding preferences of RBPs have been noticed in previous studies (Lebedeva et al. 2011; Wagnon et al. 2012). However, the link between positional biases of RBP binding sites and their functional consequences has not been well established. PRAS reveals that the distal end of 3' UTR binding is predictive of CELF4-stabilized targets. The distal end bias of CELF4-stabilized targets suggests possible molecular mechanism(s) by which CELF4 regulates its mRNAs. It has been reported that poly(A) tails enhance the stability of mRNAs (Subtelny et al. 2014). The proximity between poly(A) tails and the distal 3' UTRs suggests possible connections with poly(A) tail functions, such as mRNA stability, polyadenylation itself or promotion of translational reinitiation – possibilities to be explored in future experimental studies. CELF1 is known to recruit cytoplasmic deadenylases (Vlasova-St Louis et al. 2013) and the extent of mRNA degradation is positively correlated to CELF1's binding magnitude to the 3' UTRs (Wang et al. 2015). Based on the finding in the previous study (Wang et al. 2015) that CELF1 binding is enriched in the 3' end of the 3' UTR, we further found that this binding bias shows strong predictive ability to CELF1-degraded targets (Figure 3.3 b). We also demonstrated the potential of PRAS in the large-scale

applications by showing the better performance of PRAS than other methods in identifying the targets of RNA decay related RBPs from ENCODE (Consortium 2012). These results again strengthen the relationship between the regulatory functions of the RBPs and their binding positions.

3.4 AVAILABILITY AND FUTURE DIRECTIONS

PRAS is implemented in Python and R and is freely available at https://github.com/ouyang-lab/PRAS. PRAS can be applied widely to identify the functional targets of any RBPs with CLIP-seq peaks. For RBPs with a known post-transcriptional function, the functional targets may be identified with a corresponding reference position that is related to that function (e.g. splicing sites for alternative splicing). PRAS can also be combined with other types of information, such as sequence motifs, conservation, and perturbation data to predict RBP functional targets using integrative approaches such as (Zhang et al. 2010). In addition, future versions of PRAS can be extended to study the co-regulations of multiple RBPs by being applied to a set of interested RBPs simultaneously and evaluating the importance of different reference sites on the targets.

CHAPTER 4: C-MYC REGULATES MRNA TRANSLATION EFFICIENCY AND START-SITE SELECTION IN LYMPHOMA

4.1 INTRODUCTION

MYC, known as an oncogenic transcription factor, has been reported to play an essential role in both normal and malignant cell bionomy. Existing studies focus on the transcriptional effect of MYC, and it is known that consensus E-boxes are the immediate targets of MYC to achieve selective effects (Land et al. 1983; Blackwood and Eisenman 1991). In the recent studies of MYC biology, the augmentation function of the expression of active genes has been reported, so MYC has also been described as a "global amplifier" (Lin et al. 2012; Nie et al. 2012). More recently, the interactions between MYC and coactivators or inhibitors are reported to contribute to the transcriptional effects, which refines the understanding of its global-amplification function (Ouyang et al. 2009; Walz et al. 2014; Kress et al. 2015). Besides the role of transcription factor, it has also been reported that MYC is associated with the control of mRNA translation. However, these reported effects of MYC are secondary, such as changes in the expression of ribosomal proteins and translation factors, or mRNA capping (Schlosser et al. 2003; Arabi et al. 2005; Grandori et al. 2005; Cole and Cowling 2009; van Riggelen et al. 2010).

The translational regulation in cancer have been reported in metabolism, migration, and metastasis (Topisirovic and Sonenberg 2011; Hsieh et al. 2012; Pourdehnad et al. 2013; Elkon et al. 2015; Truitt et al. 2015; Lindqvist et al. 2018), which have been largely attributed to activation of the mTOR/4EBP1/eIF4E signaling axis (Lin et al. 2008; Hardie et al. 2012; Bhat et al. 2015; Morita et al. 2017; Saxton and Sabatini 2017). In the present study, we show the effect of MYC on global mRNA translation efficiency (TE), and on translation start-site usage using harringtonine to arrest the initiating ribosomes.

4.2 RESULTS

4.2.1 MYC has global and specific effects on mRNA translation in lymphoma cells

We performed transcriptome-scale ribosome profiling to identify precisely which mRNAs are affected translationally by MYC. This method isolates changes in translation from changes in transcription

by relating ribosome-protected fragment (RF) reads to total mRNA levels (Pajic et al. 2000; Ingolia et al. 2009; Wolfe et al. 2014). We performed the study in triplicates on P493-6 human B lymphoma cells that express MYC in a doxycycline-sensitive manner and compared high- and low-MYC states at a 24-h time point. Important quality-control data are shown in Supplementary Figure S4.1 a–c, and described in the figure legend. For most transcripts, the change in translation as indicated by ribosome coverage (RF reads) was proportional to the change in mRNA abundance (r = 0.41; Figure 4.1, indicated by the gray diagonal line). However, using a strict statistical cutoff at q < 0.01, we identified mRNAs whose translation was disproportionally affected by MYC. Specifically, the TE was up-regulated in the high-MYC state for 882 mRNAs (TE up) and decreased (TE down) for 315 mRNAs (Figure 4.1 and Supplementary Figure S4.1 d, marked in red and blue, respectively).

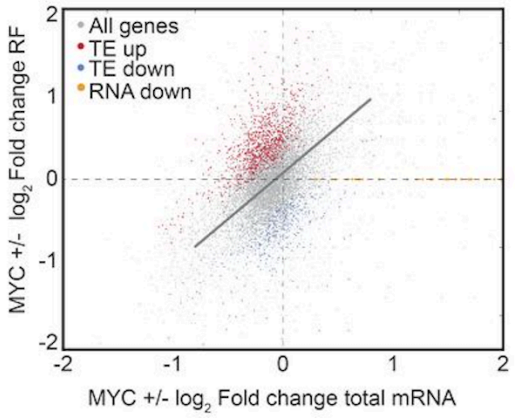


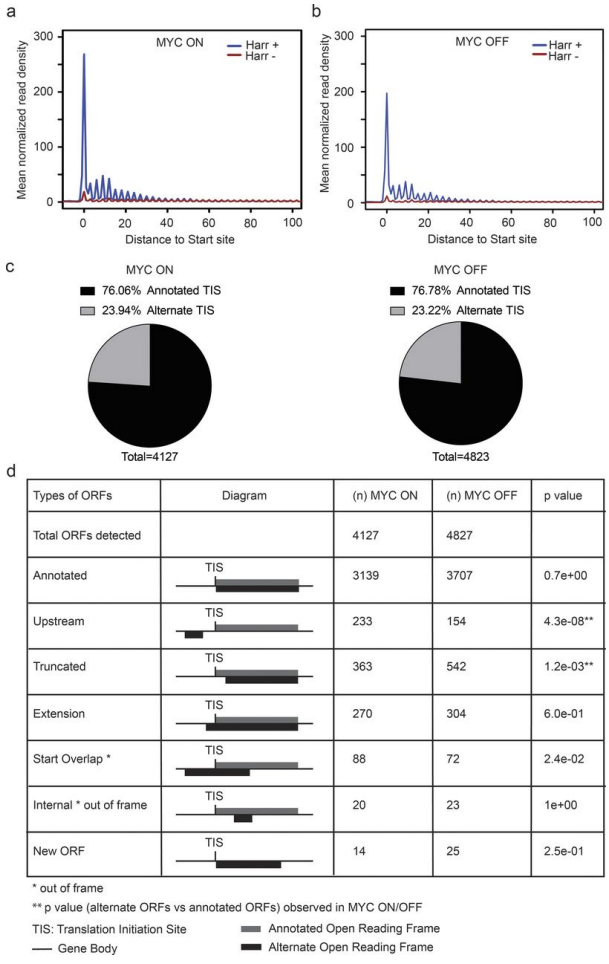
Figure 4.1 Change in total mRNA levels versus change in RF reads in the presence or absence of MYC in P493-6 cells. The linear function indicates proportional changes in both: genes with a significantly (q < 0.01) disproportional increase in TE (TE up, red) or decrease (TE down, blue). Ribosome footprinting was performed in three biological replicates for each group.

4.2.2 MYC changes translation initiation sites (TISs) and open reading frames (ORFs)

Next, we wanted to explore to what extent MYC affects translation start sites and potentially the integrity of ORFs. We experimentally mapped TISs in the presence and absence of MYC by performing ribosome profiling in the presence of harringtonine. Briefly, harringtonine arrests the initiating ribosomes, and this is readily detectable as an RNase I–protected sequence peak overlaying the actual start site (Fresno et al. 1977; Robert et al. 2009). We performed the experiment in triplicates, removed one outlier sample from further analyses (Supplementary Figure S4.2 a), and discarded irrelevant reads. A metagene analysis (for positions −2 to +90) confirmed a robust harringtonine-induced arrest (Figure 4.2 a and b). Briefly, we determined the peptidyl (P-site) offset for different read lengths by aligning the ribosome-protected reads to the annotated AUG start codons (Supplementary Figure S4.2 b). On average the P-site offset was 12 nucleotides, and we used this number to identify alternate TISs (ATISs; Supplementary Figure S4.2 c). In both conditions, most transcripts initiated from a single TIS (Supplementary Figure S4.2 d and e). We used the ORF-RATER algorithm to identify all consensus and variant TIS in each condition; the program identifies TISs based on a ribosome-protected RNA sequence peak and the presence of a potential start site NUG, where N represent A/T/G/C. Briefly, we grouped annotated RNA isoforms that share a genomic position on the same strand into "transcript families." We used an ORF-RATER score > 0.8 as a significant cutoff and used only these ORFs for further analyses (Fields et al. 2015).

We noticed a surprising variation in actual versus predicted TISs. The predicted TIS was the first consensus AUG start codon and gave rise to a functional ORF and protein; the ATIS reflects actual ribosome accumulation upon initiation arrest with harringtonine. Overall, we detected ~23% of ATISs in both MYC conditions (Figure 4.2 c). Generally, in the presence of MYC, we detected a significant ($P = 4.3 \times 10{-}08$) increase in the usage of 5′ upstream ATISs that corresponded to upstream ORFs (uORFs) and new ORFs that overlapped with the annotated start sites (Figure 4.2 d and Supplementary Figure S4.2 f). This change was also detected by increased 80S ribosome coverage across 5′ upstream mRNA sequences (Supplementary Figure S4.2 g). Conversely, as a general rule with some exceptions, we saw a significant

(P = 1.2 × 10−03) shift to an ATIS downstream (3′) from the annotated site in the absence of MYC; the latter is expected to give rise to N-terminal truncations (Figure 4.2 d and Supplementary Figure S4.2 f). This change was also reflected in the start codon choice, and alternate ORFs typically initiated from near cognate CUG, GUG, or UUG codons instead of the annotated AUG codon (P < 0.05; Supplementary Figure S4.2 h and i). Hence, high- and low-MYC conditions lead to surprising usage of up- and downstream ATIS, respectively.



Types of ORFs	Diagram	(n) MYC ON	(n) MYC OFF	p value
Total ORFs detected		4127	4827	
Annotated	TIS	3139	3707	0.7e+00
Upstream	TIS	233	154	4.3e-08**
Truncated	TIS	363	542	1.2e-03**
Extension	TIS	270	304	6.0e-01
Start Overlap *	TIS	88	72	2.4e-02
Internal * out of frame	TIS	20	23	1e+00
New ORF	TIS	14	25	2.5e-01

Figure 4.2 MYC affects TIS choice. (a and b) Metagene analysis of TIS detection in the presence and absence of MYC under harringtonine (Harr)-induced translation arrest (2 µg/ml; 2 min). Ribosome densities were averaged after aligning the gene density profile at the TIS to obtain the mean normalized read density. n = 3 biological replicates in each group. (c) Annotated and ATIS in all ORFs detected in MYC ON and OFF samples. n = 3 biological replicates in each group. (d) Annotated and alternate ORFs detected in the presence/absence of MYC; significance by Fisher's exact test. n = 3 biological replicates in each group.

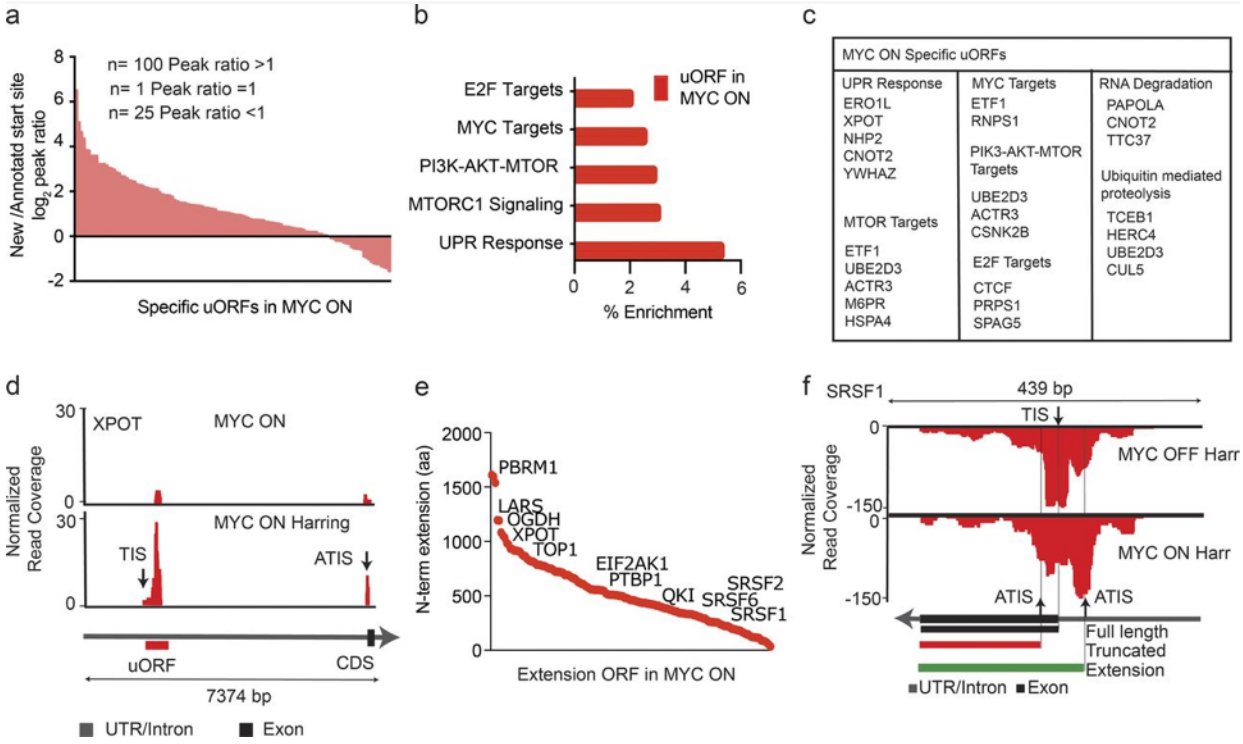


Figure 4.3 High-MYC conditions favor upstream translation initiation. (a) The peak height ratio of RF reads across the annotated TIS versus the uORF TIS indicates preferential uORF initiation for most uORF-containing genes in the MYC ON condition. n = 3 biological replicates in each group. (b) GO identifies categories of genes with MYC-activated uORFs. n = 3 biological replicates in each group. (c) List of genes harboring uORFs in the MYC ON state by KEGG category. n = 3 biological replicates in each group. (d) RF distribution with and without harringtonine for XPOT indicates uORF usage in the MYC ON state; black and red arrows indicate the predicted (TIS) and ATIS, respectively. n = 3 biological replicates in each group. (e) MYC-induced 5' extended ORFs ranked by the number of additional N-terminal amino acids. n = 3 biological replicates in each group. (f) RF read distribution across the SRSF1 transcript in high and low MYC indicates variable ATIS usage. Harr indicates harringtonine arrest. Black arrows indicate predicted TIS and ATIS, respectively. Exons shown as black squares. n = 3 biological replicates in each group.

4.2.3 Variant ORFs result in abnormal proteins

First, we examined ATIS usage in high-MYC conditions. Read count ratios from the annotated TIS and ATIS indicated that, when both were present in a transcript, the ATISs were preferred (80%) over the annotated TIS (20%; Figure 4.3 a). We do not know the biological relevance of these extended variant ORFs (ntotal = 233, nspecific = 157; Figure 4.2 d, Supplementary Figure S4.2 f). However, GO analysis indicated a significant enrichment for genes that were also MYC and E2F transcriptional targets (Figure 4.3 b and c). Recent studies indicate that under stress conditions uORFs enhance the translation of the downstream ORF (Vattem and Wek 2004; Sendoel et al. 2017). Consistently, we observed that the presence of a uORF is linked to increased or unchanged, but never to reduced, TE (Supplementary Figure S4.2 j). One example is the XPOT gene encoding exportin-T, a nuclear exporter of aminoacylated transfer RNAs, which gains a prominent uORF in high-MYC states (Arts et al. 1998)(Figure 4.3 d). In other instances, MYC activation leads to 5′ extended ORFs that encode N-terminally extended proteins. This group includes many RNA-binding proteins including SRSF family members (Figure 4.3 e). The SRSF1 transcript showed loss of initiation from the annotated TIS and relative increases of usage of both 5′ and 3′ ATIS (Figure 4.3 f). The variable use of ATISs likely affected production of the functional protein.

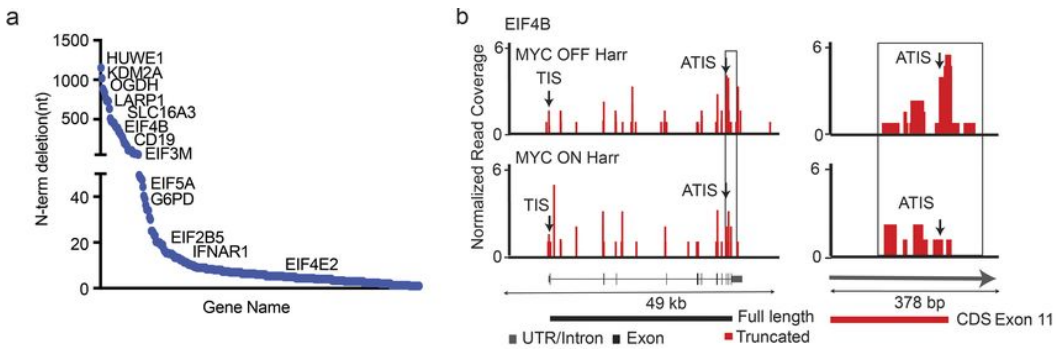


Figure 4.4 Low-MYC expression favors downstream start sites that lead to functional N-terminal truncations. (a) Genes ranked by the distance of ATIS from TIS resulting in truncated ORFs. n = 3 biological replicates in each group. (b) RF distribution across the eIF4B transcript under harringtonine (Harr) treatment in MYC ON and OFF states. Arrows indicate the predicted TIS and the ATIS, the zoomed-in region on the right illustrates differential usage of the ATIS in MYC ON and OFF states. n = 3 biological replicates in each group.

The absence of MYC favored initiation from an ATIS downstream (3′), and this would cause N-terminal truncations (Figure 4.2 d and Supplementary Figure S4.2 f). Generally, the ATIS in low-MYC conditions were typically located between 3 and 5,038 nucleotides downstream from the annotated AUG, they were typically CUG or GUG codons, and in low-MYC states, the ATIS was preferred (70%) over the annotated TIS (30%; Supplementary Figure S4.3, a–c). Some examples of N-terminally truncating start sites include mTOR and translation regulators such as LARP1, eIF4B, eIF3M, eIF2B5, eIF4E2, and eIF4B (Figure 4.4 a). The ORF-RATER identified the new ATIS based on an RF peak, and the presence of a potential start site; for example, eIF4B acquired a new ATIS in exon 11 resulting in a truncated protein that loses all relevant RNA-binding domains (Shahbazian et al. 2010)(Figure 4.4 b).

4.3 MATERIALS AND METHODS

4.3.1 ORF analysis using harringtonine data

We predicted the annotated and alternative ORFs for MYC ON and OFF samples using ORF-RATER (Fields et al. 2015). Both untreated and harringtonine-treated samples were used to perform the prediction analysis, in which harringtonine ON samples contributed to the prediction of the TIS, and harringtonine OFF samples contributed to the prediction of both the TIS and translation termination sites. We extracted the reliable uORFs, truncated ORFs, extension ORFs, start-overlap ORFs, internal ORFs, new ORFs, and annotated ORFs using 0.8 as the score cutoff suggested by Fields et al. (Fields et al. 2015). MYC ON– and MYC OFF–specific ORFs are those ORFs that were detected only in MYC ON or MYC OFF samples, respectively.

4.3.2 Alternative and annotated initiation site peak ratio analysis

For genes with ATISs, we compared the relative translation level of the alternative ORF and its corresponding annotated ORF using the peak ratio of the translation start site. We extracted the sum of the footprint read counts in the region of −30 nt to +30 nt relative to the ATIS and annotated TIS in the

harringtonine-treated samples. We calculated the peak ratio of an alternative ORF as the reads ratio of the ATIS to the annotated TIS.

4.3.3 Metagene analysis

For each gene, we calculated the mean ribosome footprint density across the positions on the longest transcript of a gene with ≥64 footprint read counts. We normalized the positional footprint density of each gene by the average footprint density. Then we scaled both the 5′UTRs and CDSs of genes to an equal number of windows and calculated the averaged signals across all genes as the metagene profile. We plotted the final metagene profile by averaging the metagene profiles across replicates.

4.3.4 CDS ribosome pause site analysis

Similarly, we used the normalized footprint density of the longest transcript of each gene that has ≥64 ribosome footprint read counts for the CDS ribosome pause-site analysis. We defined the normalized codon density as the sum of the normalized footprint density of the nucleotides at positions −1, 0, and +1 relative to the first nucleotide of that codon. We considered a codon as a ribosome pause if the codon density was ≥150. This cutoff was decided by the 0.1% quantile of the normalized codon density distribution along the CDSs. CDS ribosome pause sites were excluded if they are within 5 codons to the TIS or the translation stop sites. We aligned the metagene profile of the flanking region around the ribosome pause sites and plotted the averaged signals.

4.3.5 CDS ribosome pause-site motif analysis

We extracted the peptide sequences of the flanking region around the pause sites as the positive sequence set for the motif analysis. We choose the 5-peptide upstream and downstream flanking region to the pause site, which is 11 peptides in total including the pause-site peptide. We generated a set of random regions with the same size outside the flanking regions of the pause sites and extracted the peptide

sequences of these random regions as the negative sequence set. Then, we did the motif analysis based on the positive and negative sequence sets.

4.4 DISCUSSION

Our findings provide new insight into the biological activity of MYC and its effect on mRNA translation. We mapped the global and gene-selective effects of MYC activation on mRNA translation. MYC has surprising and profound effects on the choice of translation start sites in mammalian cells. We found that TIS choice affects the integrity of ORFs and, therefore, the production of functional proteins. In general, we observed that MYC activation shifts TIS upstream from the annotated site, whereas low-MYC states correspond to the opposite effect. The 5′ shift activates uORFs and produces overlapping and 5′ extended ORFs, whose biological activities are not yet known. The 5′ truncation seen in low-MYC conditions deletes important functional domains and may also affect protein stability. For example, the eIF4B translation initiation factor loses all of its RNA-binding domains and is unlikely to retain activity. Immediately relevant to lymphoma therapy is the effect on the CD19 cell surface receptor. Under low-MYC conditions, we observed predominant usage of a downstream TIS that results in loss of all receptor ectodomains. This change impairs detection by antibodies against the CD19 N-terminus, and it also protects lymphoma cells from attack by CD19-directed CAR-T cells. Clinically, loss of surface CD19 has been linked to resistance to CAR-T cell therapy for lymphoma although the molecular mechanism has not been defined (Sotillo et al. 2015; Perna and Sadelain 2016). Our results indicate that alternate translation start-site choice can lead to the expression of abnormal cell surface receptors. Altogether, we found physiologically relevant effects of MYC levels on the efficiency of mRNA translation and the integrity of ORFs and proteins.

CHAPTER 5: CONCLUSIONS

DNA methylation is an important epigenetic modification that undergoes dynamic changes in mammalian embryogenesis, during which both parental genomes are reprogrammed. Despite the many immunostaining studies that have assessed global methylation, the gene-specific DNA methylation patterns in bovine preimplantation embryos are unknown. Using reduced representation bisulfite sequencing, we determined genome-scale DNA methylation of bovine sperm and individual in vivo developed oocytes and preimplantation embryos. We show that (1) the major wave of genome-wide demethylation was completed by the 8-cell stage; (2) promoter methylation was significantly and inversely correlated with gene expression at the 8-cell and blastocyst stages; (3) sperm and oocytes have numerous differentially methylated regions (DMRs)—DMRs specific for sperm were strongly enriched in long terminal repeats and rapidly lost methylation in embryos; while the oocyte-specific DMRs were more frequently localized in exons and CpG islands (CGIs) and demethylated gradually across cleavage stages; (4) DMRs were also found between in vivo and in vitro matured oocytes; and (5) differential methylation between bovine gametes was confirmed in some but not all known imprinted genes. Our data provide insights into the complex epigenetic reprogramming of bovine early embryos, which serve as an important model for human preimplantation development.

RNA-protein interaction plays important roles in post-transcriptional regulation. Recent advancements in cross-linking and immunoprecipitation followed by sequencing (CLIP-seq) technologies make it possible to detect the binding peaks of a given RNA binding protein (RBP) at transcriptome scale. However, it is still challenging to predict the functional consequences of RBP binding peaks. In this study, we propose the Protein-RNA Association Strength (PRAS), which integrates the intensities and positions of the binding peaks of RBPs for functional mRNA targets prediction. We illustrate the superiority of PRAS over existing approaches on predicting the functional targets of two related but divergent CELF (CUGBP, ELAV-like factor) RBPs in mouse brain and muscle. We also demonstrate the potential of PRAS for wide adoption by applying it to the enhanced CLIP-seq (eCLIP) datasets of 37 RNA decay related RBPs in two

human cell lines. PRAS can be utilized to investigate any RBPs with available CLIP-seq peaks. PRAS is freely available at http://ouyanglab.jax.org/pras/.

The oncogenic c-MYC (MYC) transcription factor has broad effects on gene expression and cell behavior. We show that MYC alters the efficiency and quality of mRNA translation into functional proteins. Specifically, MYC drives the translation of most protein components of the electron transport chain in lymphoma cells, and many of these effects are independent from proliferation. Specific interactions of MYC-sensitive RNA-binding proteins (e.g., SRSF1/RBM42) with 5′UTR sequence motifs mediate many of these changes. Moreover, we observe a striking shift in translation initiation site usage. For example, in low-MYC conditions, lymphoma cells initiate translation of the CD19 mRNA from a site in exon 5. This results in the truncation of all extracellular CD19 domains and facilitates escape from CD19-directed CAR-T cell therapy. Together, our findings reveal MYC effects on the translation of key metabolic enzymes and immune receptors in lymphoma cells.

REFERENCE:

Adelson DL, Raison JM, Edgar RC. 2009. Characterization and distribution of retrotransposons and simple sequence repeats in the bovine genome. *Proc Natl Acad Sci U S A* **106**: 12855-12860.

Althammer S, Gonzalez-Vallinas J, Ballare C, Beato M, Eyras E. 2011. Pyicos: a versatile toolkit for the analysis of high-throughput sequencing data. *Bioinformatics* **27**: 3333-3340.

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol* **11**: R106.

Arabi A, Wu S, Ridderstrale K, Bierhoff H, Shiue C, Fatyol K, Fahlen S, Hydbring P, Soderberg O, Grummt I et al. 2005. c-Myc associates with ribosomal DNA and activates RNA polymerase I transcription. *Nat Cell Biol* **7**: 303-310.

Arts GJ, Fornerod M, Mattaj IW. 1998. Identification of a nuclear export receptor for tRNA. *Curr Biol* **8**: 305-314.

Bakhtari A, Ross PJ. 2014. DPPA3 prevents cytosine hydroxymethylation of the maternal pronucleus and is required for normal development in bovine embryos. *Epigenetics* **9**: 1271-1279.

Barlow DP, Bartolomei MS. 2014. Genomic imprinting in mammals. *Cold Spring Harb Perspect Biol* **6**.

Barone R, Fichera M, De Grandi M, Battaglia M, Lo Faro V, Mattina T, Rizzo R. 2017. Familial 18q12.2 deletion supports the role of RNA-binding protein CELF4 in autism spectrum disorders. *American journal of medical genetics Part A* **173**: 1649-1655.

Barres R, Kirchner H, Rasmussen M, Yan J, Kantor FR, Krook A, Naslund E, Zierath JR. 2013. Weight loss after gastric bypass surgery in human obesity remodels promoter methylation. *Cell Rep* **3**: 1020-1027.

Barres R, Osler ME, Yan J, Rune A, Fritz T, Caidahl K, Krook A, Zierath JR. 2009. Non-CpG methylation of the PGC-1alpha promoter through DNMT3B controls mitochondrial density. *Cell Metab* **10**: 189-198.

Bartolomei MS. 2009. Genomic imprinting: employing and avoiding epigenetic processes. *Genes Dev* **23**: 2124-2133.

Bartolomei MS, Tilghman SM. 1997. Genomic imprinting in mammals. *Annu Rev Genet* **31**: 493-525.

Bhat M, Robichaud N, Hulea L, Sonenberg N, Pelletier J, Topisirovic I. 2015. Targeting the translation machinery in cancer. *Nat Rev Drug Discov* **14**: 261-278.

Bird A. 2002. DNA methylation patterns and epigenetic memory. *Genes Dev* **16**: 6-21.

Blackwood EM, Eisenman RN. 1991. Max: a helix-loop-helix zipper protein that forms a sequence-specific DNA-binding complex with Myc. *Science* **251**: 1211-1217.

Braude P, Bolton V, Moore S. 1988. Human gene expression first occurs between the four- and eight-cell stages of preimplantation development. *Nature* **332**: 459-461.

Canovas S, Ross PJ, Kelsey G, Coy P. 2017. DNA Methylation in Embryo Development: Epigenetic Impact of ART (Assisted Reproductive Technologies). *Bioessays* **39**.

Chen B, Yun J, Kim MS, Mendell JT, Xie Y. 2014. PIPE-CLIP: a comprehensive online tool for CLIP-seq data analysis. *Genome Biol* **15**: R18.

Chen Z, Hagen DE, Elsik CG, Ji T, Morris CJ, Moon LE, Rivera RM. 2015. Characterization of global loss of imprinting in fetal overgrowth syndrome induced by assisted reproduction. *Proc Natl Acad Sci U S A* **112**: 4618-4623.

Chen Z, Robbins KM, Wells KD, Rivera RM. 2013. Large offspring syndrome: a bovine model for the human loss-of-imprinting overgrowth syndrome Beckwith-Wiedemann. *Epigenetics* **8**: 591-601.

Chenard CA, Richard S. 2008. New implications for the QUAKING RNA binding protein in human disease. *J Neurosci Res* **86**: 233-242.

Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G et al. 2008. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods* **5**: 613-619.

Cole MD, Cowling VH. 2009. Specific regulation of mRNA cap methylation by the c-Myc and E2F1 transcription factors. *Oncogene* **28**: 1169-1175.

Comoglio F, Sievers C, Paro R. 2015. Sensitive and highly resolved identification of RNA-protein interaction sites in PAR-CLIP data. *BMC Bioinformatics* **16**: 32.

Consortium EP. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57-74.

Cook KB, Hughes TR, Morris QD. 2015. High-throughput characterization of protein-RNA interactions. *Brief Funct Genomics* **14**: 74-89.

Corcoran DL, Georgiev S, Mukherjee N, Gottwein E, Skalsky RL, Keene JD, Ohler U. 2011. PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome Biol* **12**: R79.

Dasgupta T, Ladd AN. 2012. The importance of CELF control: molecular and biological roles of the CUG-BP, Elav-like family of RNA-binding proteins. *Wiley Interdiscip Rev RNA* **3**: 104-121.

Dean W, Santos F, Stojkovic M, Zakhartchenko V, Walter J, Wolf E, Reik W. 2001. Conservation of methylation reprogramming in mammalian development: aberrant reprogramming in cloned embryos. *Proc Natl Acad Sci U S A* **98**: 13734-13738.

Dobbs KB, Rodriguez M, Sudano MJ, Ortega MS, Hansen PJ. 2013. Dynamics of DNA methylation during early development of the preimplantation bovine embryo. *PLoS One* **8**: e66230.

Elkon R, Loayza-Puch F, Korkmaz G, Lopes R, van Breugel PC, Bleijerveld OB, Altelaar AF, Wolf E, Lorenzin F, Eilers M et al. 2015. Myc coordinates transcription and translation to enhance transformation and suppress invasiveness. *EMBO Rep* **16**: 1723-1736.

Erhard F, Dolken L, Jaskiewicz L, Zimmer R. 2013. PARma: identification of microRNA target sites in AGO-PAR-CLIP data. *Genome Biol* **14**: R79.

Fenger-Gron M, Fillman C, Norrild B, Lykke-Andersen J. 2005. Multiple processing body factors and the ARE binding protein TTP activate mRNA decapping. *Mol Cell* **20**: 905-915.

Fernandez-Gonzalez R, Ramirez MA, Pericuesta E, Calle A, Gutierrez-Adan A. 2010. Histone modifications at the blastocyst Axin1(Fu) locus mark the heritability of in vitro culture-induced epigenetic alterations in mice. *Biol Reprod* **83**: 720-727.

Fields AP, Rodriguez EH, Jovanovic M, Stern-Ginossar N, Haas BJ, Mertins P, Raychowdhury R, Hacohen N, Carr SA, Ingolia NT et al. 2015. A Regression-Based Analysis of Ribosome-Profiling Data Reveals a Conserved Complexity to Mammalian Translation. *Mol Cell* **60**: 816-827.

Finotello F, Di Camillo B. 2015. Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. *Brief Funct Genomics* **14**: 130-142.

Fresno M, Jimenez A, Vazquez D. 1977. Inhibition of translation in eukaryotic systems by harringtonine. *Eur J Biochem* **72**: 323-330.

Fulka H, Mrazek M, Tepla O, Fulka J, Jr. 2004. DNA methylation pattern in human zygotes and developing embryos. *Reproduction* **128**: 703-708.

Gao F, Niu Y, Sun YE, Lu H, Chen Y, Li S, Kang Y, Luo Y, Si C, Yu J et al. 2017. De novo DNA methylation during monkey pre-implantation embryogenesis. *Cell Res* **27**: 526-539.

Gebert C, Wrenzycki C, Herrmann D, Groger D, Reinhardt R, Hajkova P, Lucas-Hahn A, Carnwath J, Lehrach H, Niemann H. 2006. The bovine IGF2 gene is differentially methylated in oocyte and sperm DNA. *Genomics* **88**: 222-229.

Gebert C, Wrenzycki C, Herrmann D, Groger D, Thiel J, Reinhardt R, Lehrach H, Hajkova P, Lucas-Hahn A, Carnwath JW et al. 2009. DNA methylation in the IGF2 intragenic DMR is re-established in a sex-specific manner in bovine blastocysts after somatic cloning. *Genomics* **94**: 63-69.

Gilbert C, Kristjuhan A, Winkler GS, Svejstrup JQ. 2004. Elongator interactions with nascent mRNA revealed by RNA immunoprecipitation. *Mol Cell* **14**: 457-464.
Gkountela S, Zhang KX, Shafiq TA, Liao WW, Hargan-Calvopina J, Chen PY, Clark AT. 2015. DNA Demethylation Dynamics in the Human Prenatal Germline. *Cell* **161**: 1425-1436.
Glisovic T, Bachorik JL, Yong J, Dreyfuss G. 2008. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett* **582**: 1977-1986.
Goldberg AD, Allis CD, Bernstein E. 2007. Epigenetics: a landscape takes shape. *Cell* **128**: 635-638.
Goll MG, Bestor TH. 2005. Eukaryotic cytosine methyltransferases. *Annu Rev Biochem* **74**: 481-514.
Graf A, Krebs S, Zakhartchenko V, Schwalb B, Blum H, Wolf E. 2014. Fine mapping of genome activation in bovine embryos by RNA sequencing. *Proc Natl Acad Sci U S A* **111**: 4139-4144.
Grandori C, Gomez-Roman N, Felton-Edkins ZA, Ngouenet C, Galloway DA, Eisenman RN, White RJ. 2005. c-Myc binds to human ribosomal DNA and stimulates transcription of rRNA genes by RNA polymerase I. *Nat Cell Biol* **7**: 311-318.
Guo F, Yan L, Guo H, Li L, Hu B, Zhao Y, Yong J, Hu Y, Wang X, Wei Y et al. 2015. The Transcriptome and DNA Methylome Landscapes of Human Primordial Germ Cells. *Cell* **161**: 1437-1452.
Guo H, Zhu P, Yan L, Li R, Hu B, Lian Y, Yan J, Ren X, Lin S, Li J et al. 2014. The DNA methylation landscape of human early embryos. *Nature* **511**: 606-610.
Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano M, Jr., Jungkamp AC, Munschauer M et al. 2010. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* **141**: 129-141.
Halgren C, Bache I, Bak M, Myatt MW, Anderson CM, Brondum-Nielsen K, Tommerup N. 2012. Haploinsufficiency of CELF4 at 18q12.2 is associated with developmental and behavioral disorders, seizures, eye manifestations, and obesity. *European journal of human genetics : EJHG* **20**: 1315-1319.
Hamatani T, Carter MG, Sharov AA, Ko MS. 2004. Dynamics of global gene expression changes during mouse preimplantation development. *Dev Cell* **6**: 117-131.
Hardie DG, Ross FA, Hawley SA. 2012. AMPK: a nutrient and energy sensor that maintains energy homeostasis. *Nat Rev Mol Cell Biol* **13**: 251-262.
Hsieh AC, Liu Y, Edlind MP, Ingolia NT, Janes MR, Sher A, Shi EY, Stumpf CR, Christensen C, Bonham MJ et al. 2012. The translational landscape of mTOR signalling steers cancer initiation and metastasis. *Nature* **485**: 55-61.
Hu G, McQuiston T, Bernard A, Park YD, Qiu J, Vural A, Zhang N, Waterman SR, Blewett NH, Myers TG et al. 2015. A conserved mechanism of TOR-dependent RCK-mediated mRNA degradation regulates autophagy. *Nat Cell Biol* **17**: 930-942.
Huang da W, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**: 44-57.
Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**: 218-223.
Ingolia NT, Lareau LF, Weissman JS. 2011. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**: 789-802.
Inoue A, Zhang Y. 2011. Replication-dependent loss of 5-hydroxymethylcytosine in mouse preimplantation embryos. *Science* **334**: 194.
Iqbal K, Jin SG, Pfeifer GP, Szabo PE. 2011. Reprogramming of the paternal genome upon fertilization involves genome-wide oxidation of 5-methylcytosine. *Proc Natl Acad Sci U S A* **108**: 3642-3647.
Jackson RJ, Hellen CU, Pestova TV. 2010. The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat Rev Mol Cell Biol* **11**: 113-127.
Jaenisch R, Bird A. 2003. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet* **33 Suppl**: 245-254.

Jiang Z, Dong H, Zheng X, Marjani SL, Donovan DM, Chen J, Tian XC. 2015. mRNA Levels of Imprinted Genes in Bovine In Vivo Oocytes, Embryos and Cross Species Comparisons with Humans, Mice and Pigs. *Sci Rep* **5**: 17898.

Jiang Z, Sun J, Dong H, Luo O, Zheng X, Obergfell C, Tang Y, Bi J, O'Neill R, Ruan Y et al. 2014. Transcriptional profiles of bovine in vivo pre-implantation development. *BMC Genomics* **15**: 756.

Kalsotra A, Xiao X, Ward AJ, Castle JC, Johnson JM, Burge CB, Cooper TA. 2008. A postnatal switch of CELF and MBNL proteins reprograms alternative splicing in the developing heart. *Proc Natl Acad Sci U S A* **105**: 20333-20338.

Keene JD. 2007. RNA regulons: coordination of post-transcriptional events. *Nat Rev Genet* **8**: 533-543.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Res* **12**: 996-1006.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**: R36.

Konig J, Zarnack K, Luscombe NM, Ule J. 2012. Protein-RNA interactions: new genomic technologies and perspectives. *Nat Rev Genet* **13**: 77-83.

Konig J, Zarnack K, Rot G, Curk T, Kayikci M, Zupan B, Turner DJ, Luscombe NM, Ule J. 2010. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol* **17**: 909-915.

Kress TR, Sabo A, Amati B. 2015. MYC: connecting selective transcriptional control to global RNA production. *Nat Rev Cancer* **15**: 593-607.

Krueger F, Andrews SR. 2011. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**: 1571-1572.

Kues WA, Sudheer S, Herrmann D, Carnwath JW, Havlicek V, Besenfelder U, Lehrach H, Adjaye J, Niemann H. 2008. Genome-wide expression profiling reveals distinct clusters of transcriptional regulation during bovine preimplantation development in vivo. *Proc Natl Acad Sci U S A* **105**: 19768-19773.

Kurihara Y, Kawamura Y, Uchijima Y, Amamo T, Kobayashi H, Asano T, Kurihara H. 2008. Maintenance of genomic methylation patterns during preimplantation development requires the somatic form of DNA methyltransferase 1. *Dev Biol* **313**: 335-346.

Ladd AN, Stenberg MG, Swanson MS, Cooper TA. 2005. Dynamic balance between activation and repression regulates pre-mRNA alternative splicing during heart development. *Dev Dyn* **233**: 783-793.

Land H, Parada LF, Weinberg RA. 1983. Tumorigenic conversion of primary embryo fibroblasts requires at least two cooperating oncogenes. *Nature* **304**: 596-602.

Lebedeva S, Jens M, Theil K, Schwanhausser B, Selbach M, Landthaler M, Rajewsky N. 2011. Transcriptome-wide analysis of regulatory interactions of the RNA-binding protein HuR. *Mol Cell* **43**: 340-352.

Leibfried-Rutledge ML, Critser ES, Eyestone WH, Northey DL, First NL. 1987. Development potential of bovine oocytes matured in vitro or in vivo. *Biol Reprod* **36**: 376-383.

Lev Maor G, Yearim A, Ast G. 2015. The alternative role of DNA methylation in splicing regulation. *Trends Genet* **31**: 274-280.

Li E, Beard C, Jaenisch R. 1993. Role for DNA methylation in genomic imprinting. *Nature* **366**: 362-365.

Li J, Witten DM, Johnstone IM, Tibshirani R. 2012. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics* **13**: 523-538.

Li T, Vu TH, Ulaner GA, Littman E, Ling JQ, Chen HL, Hu JF, Behr B, Giudice L, Hoffman AR. 2005. IVF results in de novo DNA methylation and histone methylation at an Igf2-H19 imprinting epigenetic switch. *Mol Hum Reprod* **11**: 631-640.

Li Y, Sasaki H. 2011. Genomic imprinting in mammals: its life cycle, molecular mechanisms and reprogramming. *Cell Res* **21**: 466-473.
Licatalosi DD, Mele A, Fak JJ, Ule J, Kayikci M, Chi SW, Clark TA, Schweitzer AC, Blume JE, Wang X et al. 2008. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* **456**: 464-469.
Lin CJ, Cencic R, Mills JR, Robert F, Pelletier J. 2008. c-Myc and eIF4F are components of a feedforward loop that links transcription and translation. *Cancer Res* **68**: 5326-5334.
Lin CY, Loven J, Rahl PB, Paranal RM, Burge CB, Bradner JE, Lee TI, Young RA. 2012. Transcriptional amplification in tumor cells with elevated c-Myc. *Cell* **151**: 56-67.
Lin Y, Li J, Shen H, Zhang L, Papasian CJ, Deng HW. 2011. Comparative studies of de novo assembly tools for next-generation sequencing technologies. *Bioinformatics* **27**: 2031-2037.
Lindqvist LM, Tandoc K, Topisirovic I, Furic L. 2018. Cross-talk between protein synthesis, energy metabolism and autophagy in cancer. *Curr Opin Genet Dev* **48**: 104-111.
Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR. 2008. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* **133**: 523-536.
Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM et al. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**: 315-322.
Lovci MT, Ghanem D, Marr H, Arnold J, Gee S, Parra M, Liang TY, Stark TJ, Gehman LT, Hoon S et al. 2013. Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nat Struct Mol Biol* **20**: 1434-1442.
Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550.
Lucifero D, Suzuki J, Bordignon V, Martel J, Vigneault C, Therrien J, Filion F, Smith LC, Trasler JM. 2006. Bovine SNRPN methylation imprint in oocytes and day 17 in vitro-produced and somatic cell nuclear transfer embryos. *Biol Reprod* **75**: 531-538.
Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* **18**: 1509-1517.
Mattern F, Herrmann D, Heinzmann J, Hadeler KG, Bernal-Ulloa SM, Haaf T, Niemann H. 2016. DNA methylation and mRNA expression of developmentally important genes in bovine oocytes collected from donors of different age categories. *Mol Reprod Dev* **83**: 802-814.
Meins M, Schlickum S, Wilhelm C, Missbach J, Yadav S, Glaser B, Grzmil M, Burfeind P, Laccone F. 2002. Identification and characterization of murine Brunol4, a new member of the elav/bruno family. *Cytogenet Genome Res* **97**: 254-260.
Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, Jaenisch R. 2005. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res* **33**: 5868-5877.
Misirlioglu M, Page GP, Sagirkaya H, Kaya A, Parrish JJ, First NL, Memili E. 2006. Dynamics of global transcriptome in bovine matured oocytes and preimplantation embryos. *Proc Natl Acad Sci U S A* **103**: 18905-18910.
Modic M, Ule J, Sibley CR. 2013. CLIPing the brain: studies of protein-RNA interactions important for neurodegenerative disorders. *Mol Cell Neurosci* **56**: 429-435.
Moore MJ, Zhang C, Gantman EC, Mele A, Darnell JC, Darnell RB. 2014. Mapping Argonaute and conventional RNA-binding protein interactions with RNA at single-nucleotide resolution using HITS-CLIP and CIMS analysis. *Nat Protoc* **9**: 263-293.
Moraes JC, Amaral ME, Picardi PK, Calegari VC, Romanatto T, Bermudez-Echeverry M, Chiavegatto S, Saad MJ, Velloso LA. 2006. Inducible-NOS but not neuronal-NOS participate in the acute effect

of TNF-alpha on hypothalamic insulin-dependent inhibition of food intake. *FEBS Lett* **580**: 4625-4631.

Morin R, Bainbridge M, Fejes A, Hirst M, Krzywinski M, Pugh T, McDonald H, Varhol R, Jones S, Marra M. 2008. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques* **45**: 81-94.

Morita M, Prudent J, Basu K, Goyon V, Katsumura S, Hulea L, Pearl D, Siddiqui N, Strack S, McGuirk S et al. 2017. mTOR Controls Mitochondrial Dynamics and Cell Survival via MTFP1. *Mol Cell* **67**: 922-935 e925.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**: 621-628.

Mukherjee N, Corcoran DL, Nusbaum JD, Reid DW, Georgiev S, Hafner M, Ascano M, Jr., Tuschl T, Ohler U, Keene JD. 2011. Integrative regulatory mapping indicates that the RNA-binding protein HuR couples pre-mRNA processing and mRNA stability. *Mol Cell* **43**: 327-339.

Mukhopadhyay D, Houchen CW, Kennedy S, Dieckgraefe BK, Anant S. 2003. Coupled mRNA stabilization and translational silencing of cyclooxygenase-2 by a novel RNA binding protein, CUGBP2. *Mol Cell* **11**: 113-126.

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**: 1344-1349.

Nakamura T, Arai Y, Umehara H, Masuhara M, Kimura T, Taniguchi H, Sekimoto T, Ikawa M, Yoneda Y, Okabe M et al. 2007. PGC7/Stella protects against DNA demethylation in early embryogenesis. *Nat Cell Biol* **9**: 64-71.

Nakamura T, Liu YJ, Nakashima H, Umehara H, Inoue K, Matoba S, Tachibana M, Ogura A, Shinkai Y, Nakano T. 2012. PGC7 binds histone H3K9me2 to protect against conversion of 5mC to 5hmC in early embryos. *Nature* **486**: 415-419.

Nie Z, Hu G, Wei G, Cui K, Yamane A, Resch W, Wang R, Green DR, Tessarollo L, Casellas R et al. 2012. c-Myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells. *Cell* **151**: 68-79.

Niwa H, Toyooka Y, Shimosato D, Strumpf D, Takahashi K, Yagi R, Rossant J. 2005. Interaction between Oct3/4 and Cdx2 determines trophectoderm differentiation. *Cell* **123**: 917-929.

O'Doherty AM, Magee DA, O'Shea LC, Forde N, Beltman ME, Mamo S, Fair T. 2015. DNA methylation dynamics at imprinted genes during bovine pre-implantation embryo development. *BMC Dev Biol* **15**: 13.

O'Doherty AM, O'Shea LC, Fair T. 2012. Bovine DNA methylation imprints are established in an oocyte size-specific manner, which are coordinated with the expression of the DNMT3 family proteins. *Biol Reprod* **86**: 67.

Okae H, Chiba H, Hiura H, Hamada H, Sato A, Utsunomiya T, Kikuchi H, Yoshida H, Tanaka A, Suyama M et al. 2014. Genome-wide analysis of DNA methylation dynamics during early human development. *PLoS Genet* **10**: e1004868.

Okano M, Bell DW, Haber DA, Li E. 1999. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99**: 247-257.

Ouyang Z, Zhou Q, Wong WH. 2009. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proc Natl Acad Sci U S A* **106**: 21521-21526.

Pajic A, Spitkovsky D, Christoph B, Kempkes B, Schuhmacher M, Staege MS, Brielmeier M, Ellwart J, Kohlhuber F, Bornkamm GW et al. 2000. Cell cycle activation by c-myc in a burkitt lymphoma model cell line. *Int J Cancer* **87**: 787-793.

Patil V, Ward RL, Hesson LB. 2014. The evidence for functional non-CpG methylation in mammalian cells. *Epigenetics* **9**: 823-828.

Perna F, Sadelain M. 2016. Myeloid leukemia switch as immune escape from CD19 chimeric antigen receptor (CAR) therapy. *Transl Cancer Res* **5**: S221-S225.
Petrussa L, Van de Velde H, De Rycke M. 2016. Similar kinetics for 5-methylcytosine and 5-hydroxymethylcytosine during human preimplantation development in vitro. *Mol Reprod Dev* **83**: 594-605.
Pourdehnad M, Truitt ML, Siddiqi IN, Ducker GS, Shokat KM, Ruggero D. 2013. Myc and mTOR converge on a common node in protein synthesis control that confers synthetic lethality in Myc-driven cancers. *Proc Natl Acad Sci U S A* **110**: 11988-11993.
Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841-842.
Reis e Silva AR, Bruno C, Fleurot R, Daniel N, Archilla C, Peynot N, Lucci CM, Beaujean N, Duranthon V. 2012. Alteration of DNA demethylation dynamics by in vitro culture conditions in rabbit pre-implantation embryos. *Epigenetics* **7**: 440-446.
Rizos D, Ward F, Duffy P, Boland MP, Lonergan P. 2002. Consequences of bovine oocyte maturation, fertilization or early embryo development in vitro versus in vivo: implications for blastocyst yield and blastocyst quality. *Mol Reprod Dev* **61**: 234-248.
Robert F, Carrier M, Rawe S, Chen S, Lowe S, Pelletier J. 2009. Altering chemosensitivity by modulating translation elongation. *PLoS One* **4**: e5428.
Robinson MD, Oshlack A. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11**: R25.
Rot G, Wang Z, Huppertz I, Modic M, Lence T, Hallegger M, Haberman N, Curk T, von Mering C, Ule J. 2017. High-Resolution RNA Maps Suggest Common Principles of Splicing and Polyadenylation Regulation by TDP-43. *Cell Rep* **19**: 1056-1067.
Ruggero D, Pandolfi PP. 2003. Does the ribosome translate cancer? *Nat Rev Cancer* **3**: 179-192.
Sakurai N, Takahashi K, Emura N, Fujii T, Hirayama H, Kageyama S, Hashizume T, Sawai K. 2016. The Necessity of OCT-4 and CDX2 for Early Development and Gene Expression Involved in Differentiation of Inner Cell Mass and Trophectoderm Lineages in Bovine Embryos. *Cell Reprogram* **18**: 309-318.
Salilew-Wondim D, Fournier E, Hoelker M, Saeed-Zidane M, Tholen E, Looft C, Neuhoff C, Besenfelder U, Havlicek V, Rings F et al. 2015. Genome-Wide DNA Methylation Patterns of Bovine Blastocysts Developed In Vivo from Embryos Completed Different Stages of Development In Vitro. *PLoS One* **10**: e0140467.
Salvaing J, Li Y, Beaujean N, O'Neill C. 2015. Determinants of valid measurements of global changes in 5'-methylcytosine and 5'-hydroxymethylcytosine by immunolocalisation in the early embryo. *Reprod Fertil Dev* **27**: 755-764.
Santos F, Hendrich B, Reik W, Dean W. 2002. Dynamic reprogramming of DNA methylation in the early mouse embryo. *Dev Biol* **241**: 172-182.
Saxton RA, Sabatini DM. 2017. mTOR Signaling in Growth, Metabolism, and Disease. *Cell* **169**: 361-371.
Schlosser I, Holzel M, Murnseer M, Burtscher H, Weidle UH, Eick D. 2003. A role for c-Myc in the regulation of ribosomal RNA processing. *Nucleic Acids Res* **31**: 6148-6156.
Seisenberger S, Andrews S, Krueger F, Arand J, Walter J, Santos F, Popp C, Thienpont B, Dean W, Reik W. 2012. The dynamics of genome-wide DNA methylation reprogramming in mouse primordial germ cells. *Mol Cell* **48**: 849-862.
Sela N, Kim E, Ast G. 2010. The role of transposable elements in the evolution of non-mammalian vertebrates and invertebrates. *Genome Biol* **11**: R59.
Sendoel A, Dunn JG, Rodriguez EH, Naik S, Gomez NC, Hurwitz B, Levorse J, Dill BD, Schramek D, Molina H et al. 2017. Translation from unconventional 5' start sites drives tumour initiation. *Nature* **541**: 494-499.

Shah A, Qian Y, Weyn-Vanhentenryck SM, Zhang C. 2017. CLIP Tool Kit (CTK): a flexible and robust pipeline to analyze CLIP sequencing data. *Bioinformatics* **33**: 566-567.

Shahbazian D, Parsyan A, Petroulakis E, Hershey J, Sonenberg N. 2010. eIF4B controls survival and proliferation and is regulated by proto-oncogenic signaling pathways. *Cell Cycle* **9**: 4106-4109.

Shirane K, Toh H, Kobayashi H, Miura F, Chiba H, Ito T, Kono T, Sasaki H. 2013. Mouse oocyte methylomes at base resolution reveal genome-wide accumulation of non-CpG methylation and role of DNA methyltransferases. *PLoS Genet* **9**: e1003439.

Smallwood SA, Tomizawa S, Krueger F, Ruf N, Carli N, Segonds-Pichon A, Sato S, Hata K, Andrews SR, Kelsey G. 2011. Dynamic CpG island methylation landscape in oocytes and preimplantation embryos. *Nat Genet* **43**: 811-814.

Smith SL, Everts RE, Sung LY, Du F, Page RL, Henderson B, Rodriguez-Zas SL, Nedambale TL, Renard JP, Lewin HA et al. 2009. Gene expression profiling of single bovine embryos uncovers significant effects of in vitro maturation, fertilization and culture. *Mol Reprod Dev* **76**: 38-47.

Smith ZD, Chan MM, Humm KC, Karnik R, Mekhoubad S, Regev A, Eggan K, Meissner A. 2014. DNA methylation dynamics of the human preimplantation embryo. *Nature* **511**: 611-615.

Smith ZD, Chan MM, Mikkelsen TS, Gu H, Gnirke A, Regev A, Meissner A. 2012. A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nature* **484**: 339-344.

Smith ZD, Meissner A. 2013. DNA methylation: roles in mammalian development. *Nat Rev Genet* **14**: 204-220.

Sotillo E, Barrett DM, Black KL, Bagashev A, Oldridge D, Wu G, Sussman R, Lanauze C, Ruella M, Gazzara MR et al. 2015. Convergence of Acquired Mutations and Alternative Splicing of CD19 Enables Resistance to CART-19 Immunotherapy. *Cancer Discov* **5**: 1282-1295.

Strumpf D, Mao CA, Yamanaka Y, Ralston A, Chawengsaksophak K, Beck F, Rossant J. 2005. Cdx2 is required for correct cell fate specification and differentiation of trophectoderm in the mouse blastocyst. *Development* **132**: 2093-2102.

Subramaniam D, Natarajan G, Ramalingam S, Ramachandran I, May R, Queimado L, Houchen CW, Anant S. 2008. Translation inhibition during cell cycle arrest and apoptosis: Mcl-1 is a novel target for RNA binding protein CUGBP2. *Am J Physiol Gastrointest Liver Physiol* **294**: G1025-1032.

Subtelny AO, Eichhorn SW, Chen GR, Sive H, Bartel DP. 2014. Poly(A)-tail profiling reveals an embryonic switch in translational control. *Nature* **508**: 66-71.

Sutcliffe AG, Peters CJ, Bowdin S, Temple K, Reardon W, Wilson L, Clayton-Smith J, Brueton LA, Bannister W, Maher ER. 2006. Assisted reproductive therapies and imprinting disorders--a preliminary British survey. *Hum Reprod* **21**: 1009-1011.

Tian XC. 2014. Genomic imprinting in farm animals. *Annu Rev Anim Biosci* **2**: 23-40.

Timchenko LT, Timchenko NA, Caskey CT, Roberts R. 1996. Novel proteins with binding specificity for DNA CTG repeats and RNA CUG repeats: implications for myotonic dystrophy. *Hum Mol Genet* **5**: 115-121.

Tomizawa S, Kobayashi H, Watanabe T, Andrews S, Hata K, Kelsey G, Sasaki H. 2011. Dynamic stage-specific changes in imprinted differentially methylated regions during early mammalian development and prevalence of non-CpG methylation in oocytes. *Development* **138**: 811-820.

Topisirovic I, Sonenberg N. 2011. mRNA translation and energy metabolism in cancer: the role of the MAPK and mTORC1 pathways. *Cold Spring Harb Symp Quant Biol* **76**: 355-367.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**: 562-578.

Truitt ML, Conn CS, Shi Z, Pang X, Tokuyasu T, Coady AM, Seo Y, Barna M, Ruggero D. 2015. Differential Requirements for eIF4E Dose in Normal Development and Cancer. *Cell* **162**: 59-71.

Ule J, Jensen KB, Ruggiu M, Mele A, Ule A, Darnell RB. 2003. CLIP identifies Nova-regulated RNA networks in the brain. *Science* **302**: 1212-1215.

Uren PJ, Bahrami-Samani E, Burns SC, Qiao M, Karginov FV, Hodges E, Hannon GJ, Sanford JR, Penalva LO, Smith AD. 2012. Site identification in high-throughput RNA-protein interaction data. *Bioinformatics* **28**: 3013-3020.

Urrego R, Bernal-Ulloa SM, Chavarria NA, Herrera-Puerta E, Lucas-Hahn A, Herrmann D, Winkler S, Pache D, Niemann H, Rodriguez-Osorio N. 2017. Satellite DNA methylation status and expression of selected genes in Bos indicus blastocysts produced in vivo and in vitro. *Zygote* **25**: 131-140.

Van Nostrand EL, Freese P, Pratt GA, Wang X, Wei X, Xiao R, Blue SM, Chen J-Y, Cody NAL, Dominguez D et al. 2018. A Large-Scale Binding and Functional Map of Human RNA Binding Proteins. *bioRxiv* doi:10.1101/179648: 179648.

Van Nostrand EL, Pratt GA, Shishkin AA, Gelboin-Burkhart C, Fang MY, Sundararaman B, Blue SM, Nguyen TB, Surka C, Elkins K et al. 2016. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat Methods* **13**: 508-514.

van Riggelen J, Yetil A, Felsher DW. 2010. MYC as a regulator of ribosome biogenesis and protein synthesis. *Nat Rev Cancer* **10**: 301-309.

Vattem KM, Wek RC. 2004. Reinitiation involving upstream ORFs regulates ATF4 mRNA translation in mammalian cells. *Proc Natl Acad Sci U S A* **101**: 11269-11274.

Vlasova-St Louis I, Dickson AM, Bohjanen PR, Wilusz CJ. 2013. CELFish ways to modulate mRNA decay. *Biochim Biophys Acta* **1829**: 695-707.

Wagnon JL, Briese M, Sun W, Mahaffey CL, Curk T, Rot G, Ule J, Frankel WN. 2012. CELF4 regulates translation and local abundance of a vast set of mRNAs, including genes associated with regulation of synaptic function. *PLoS Genet* **8**: e1003067.

Wagnon JL, Mahaffey CL, Sun W, Yang Y, Chao HT, Frankel WN. 2011. Etiology of a genetically complex seizure disorder in Celf4 mutant mice. *Genes, brain, and behavior* **10**: 765-777.

Walz S, Lorenzin F, Morton J, Wiese KE, von Eyss B, Herold S, Rycak L, Dumay-Odelot H, Karim S, Bartkuhn M et al. 2014. Activation and repression by oncogenic MYC shape tumour-specific gene expression profiles. *Nature* **511**: 483-487.

Wan LB, Bartolomei MS. 2008. Regulation of imprinting in clusters: noncoding RNAs versus insulators. *Adv Genet* **61**: 207-223.

Wang ET, Ward AJ, Cherone JM, Giudice J, Wang TT, Treacy DJ, Lambert NJ, Freese P, Saxena T, Cooper TA et al. 2015. Antagonistic regulation of mRNA expression and splicing by CELF and MBNL proteins. *Genome Res* **25**: 858-871.

Wang L, Zhang J, Duan J, Gao X, Zhu W, Lu X, Yang L, Zhang J, Li G, Ci W et al. 2014a. Programming and inheritance of parental DNA methylomes in mammals. *Cell* **157**: 979-991.

Wang QT, Piotrowska K, Ciemerych MA, Milenkovic L, Scott MP, Davis RW, Zernicka-Goetz M. 2004. A genome-wide study of gene activity reveals developmental signaling pathways in the preimplantation mouse embryo. *Dev Cell* **6**: 133-144.

Wang T, Chen B, Kim M, Xie Y, Xiao G. 2014b. A model-based approach to identify binding sites in CLIP-Seq data. *PLoS One* **9**: e93248.

Wang T, Xie Y, Xiao G. 2014c. dCLIP: a computational approach for comparative CLIP-seq analyses. *Genome Biol* **15**: R11.

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**: 57-63.

Watson AJ. 2007. Oocyte cytoplasmic maturation: a key mediator of oocyte and embryo developmental competence. *J Anim Sci* **85**: E1-3.

Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett CJ, Rogers J, Bahler J. 2008. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453**: 1239-1243.

Wilkie GS, Dickson KS, Gray NK. 2003. Regulation of mRNA translation by 5'- and 3'-UTR-binding factors. *Trends Biochem Sci* **28**: 182-188.

Wolfe AL, Singh K, Zhong Y, Drewe P, Rajasekhar VK, Sanghvi VR, Mavrakis KJ, Jiang M, Roderick JE, Van der Meulen J et al. 2014. RNA G-quadruplexes cause eIF4A-dependent oncogene translation in cancer. *Nature* **513**: 65-70.

Wossidlo M, Nakamura T, Lepikhov K, Marques CJ, Zakhartchenko V, Boiani M, Arand J, Nakano T, Reik W, Walter J. 2011. 5-Hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming. *Nat Commun* **2**: 241.

Yan L, Yang M, Guo H, Yang L, Wu J, Li R, Liu P, Lian Y, Zheng X, Yan J et al. 2013. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat Struct Mol Biol* **20**: 1131-1139.

Yang Y, Mahaffey CL, Berube N, Maddatu TP, Cox GA, Frankel WN. 2007. Complex seizure disorder caused by Brunol4 deficiency in mice. *PLoS Genet* **3**: e124.

Yokoshi M, Li Q, Yamamoto M, Okada H, Suzuki Y, Kawahara Y. 2014. Direct binding of Ataxin-2 to distinct elements in 3' UTRs promotes mRNA stability and protein expression. *Mol Cell* **55**: 186-198.

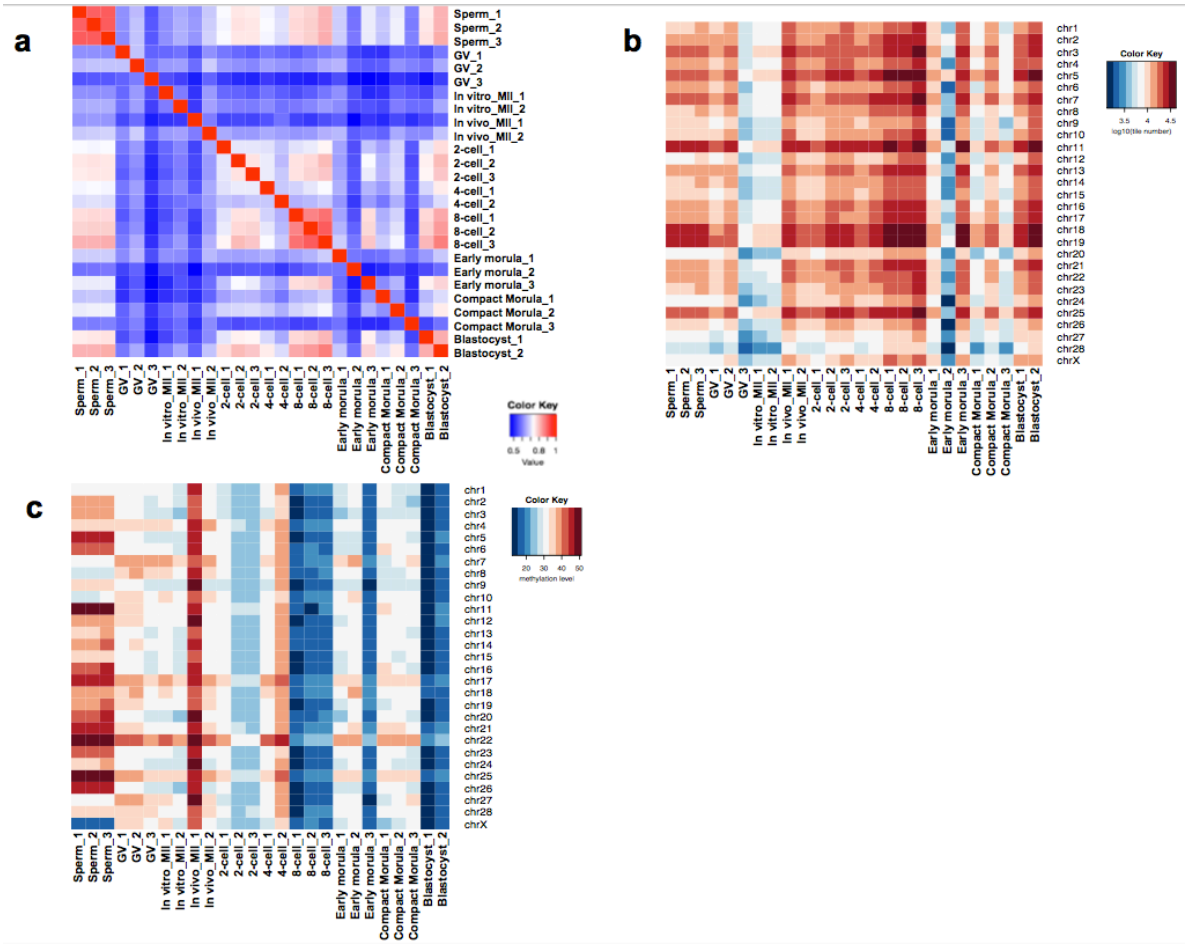
Young LE, Sinclair KD, Wilmut I. 1998. Large offspring syndrome in cattle and sheep. *Rev Reprod* **3**: 155-163.

Zhang C, Frias MA, Mele A, Ruggiu M, Eom T, Marney CB, Wang H, Licatalosi DD, Fak JJ, Darnell RB. 2010. Integrative modeling defines the Nova splicing-regulatory network and its combinatorial controls. *Science* **329**: 439-443.

Zhang S, Chen X, Wang F, An X, Tang B, Zhang X, Sun L, Li Z. 2016. Aberrant DNA methylation reprogramming in bovine SCNT preimplantation embryos. *Sci Rep* **6**: 30345.

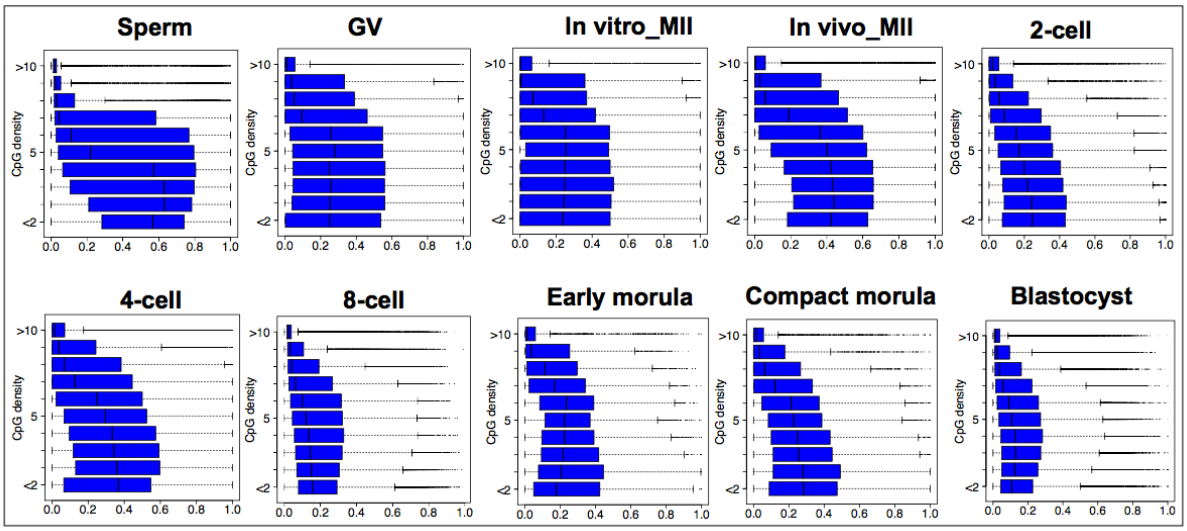
APPENDIX A: SUPPLEMENTARY FIGURES

Supplementary Figure S2.1



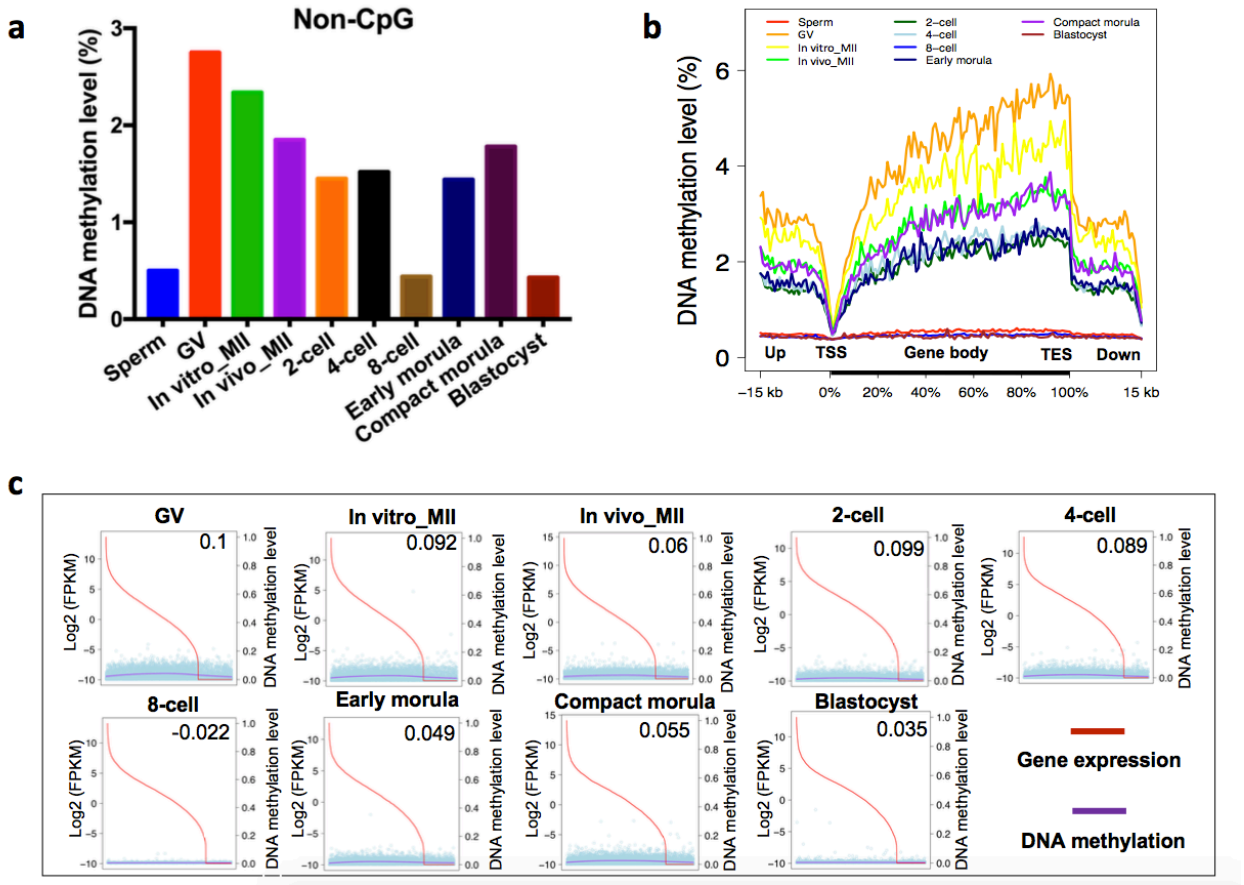
(a) Pearson correlation heatmap of DNA methylomes. GV: germinal vesicle stage oocytes; in vitro_MII: MII oocytes matured in vitro; in vivo_MII: MII oocytes matured in vivo. The numbers attached to the sample names indicate biological replicates. The color from blue to red indicates the correlation coefficient of low to high. (b) Histograms of the numbers of 100-base-pair (bp) CpG tiles captured on each chromosome across developmental stages. (c) Histograms of the average methylation levels of the 100-base-pair (bp) CpG tiles captured on each chromosome across developmental stages.

Supplementary Figure S2.2



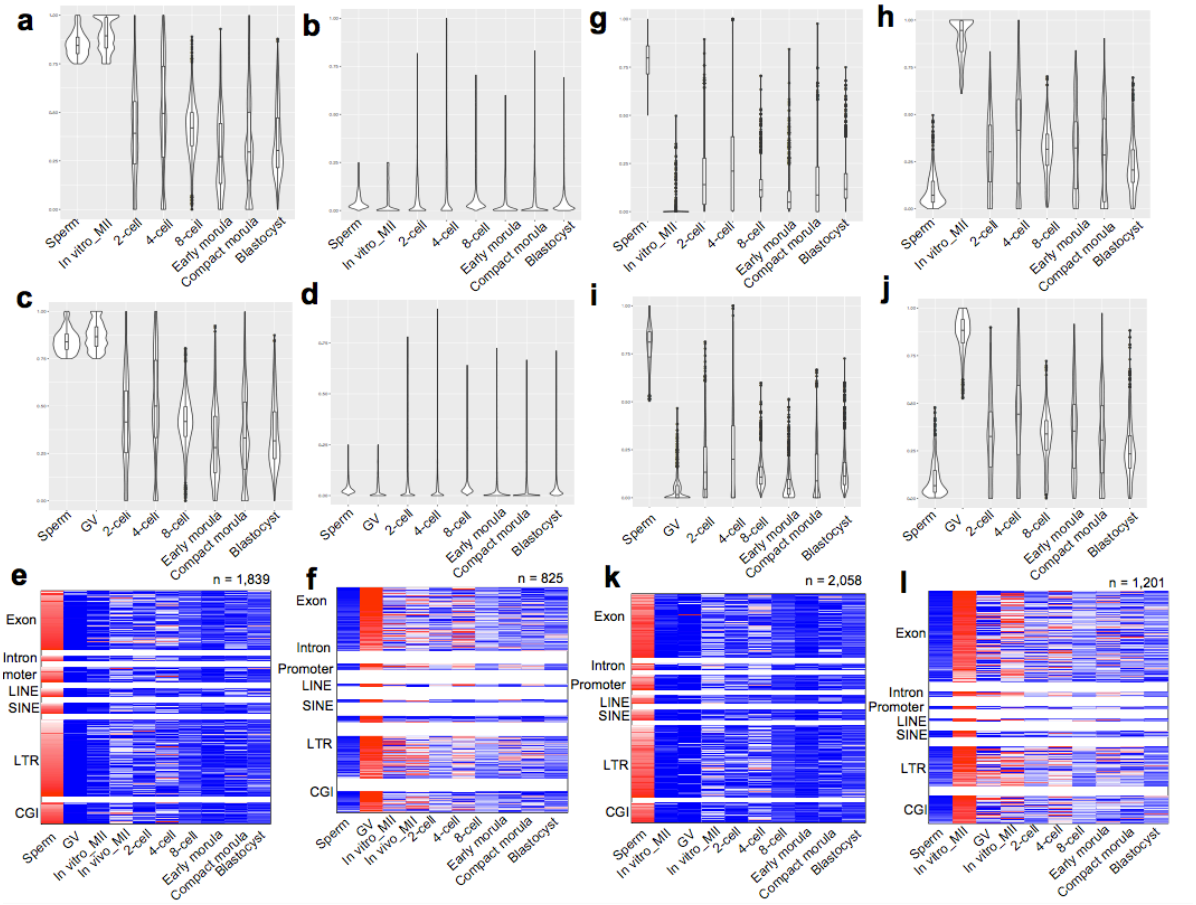
Box plots of methylation levels at each stage across local CpG densities.

Supplementary Figure S2.3



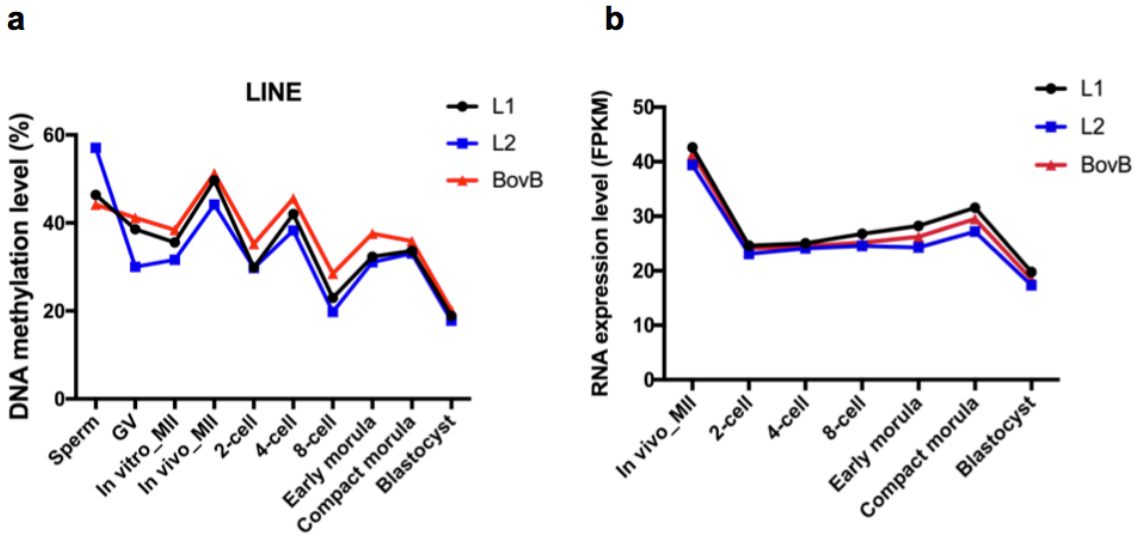
(a) The non-CpG methylation levels across each stage of bovine gametes and early embryos. The averaged non-CpG DNA methylation level of each developmental stage is calculated based on the overlapped 100-base-pair (bp) tiles detected in all of the developmental stages analyzed. (b) The averaged non-CpG DNA methylation levels along the gene bodies and 15 kilobases (kb) upstream of the transcription start sites (TSS) and 15 kb downstream of the transcription end site (TES) of all reference genes. (c) Scatter plots of non-CpG DNA methylation levels of gene body regions and the relative expression levels of corresponding genes. The log2 of the gene expression levels (FPKM) was calculated and is presented. The red and blue curves in each plot represent gene expression levels and non-CpG DNA methylation levels in gene body regions, respectively.

Supplementary Figure S2.4



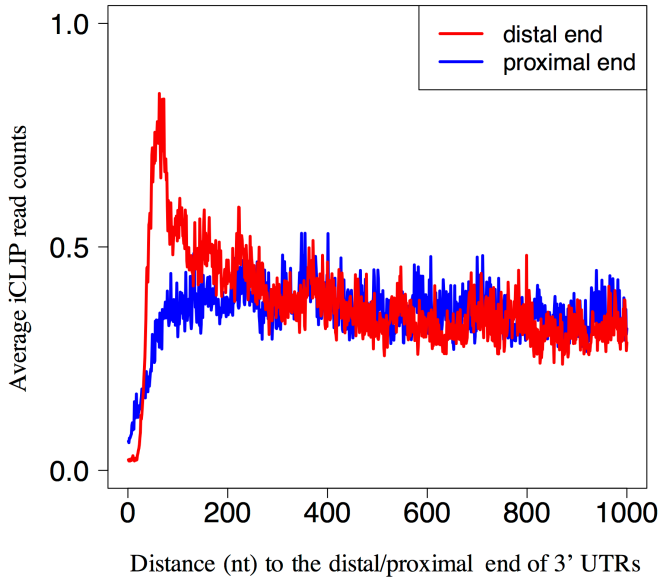
DNA methylation changes of DMRs of bovine gametes during preimplantation development. (a and b) Histogrm plots of DNA methylation levels for hypermethylated and hypomethylated 100-bp tiles in sperm and in vitro matured oocytes across early embryonic development stages. (c and d) Histogram plots of DNA methylation levels for hypermethylated and hypomethylated 100-bp tiles in sperm and GV oocytes across early embryonic development stages. (e) Heatmap of the methylation level of sperm-specific DMRs between sperm and GV oocytes among different genomic regions across different developmental stages. (f) Heatmap of the methylation level of in vivo MII oocyte-specific DMRs between sperm and GV oocytes among different genomic regions across different developmental stages. (g) Box plots of DNA methylation levels of sperm-specific DMRs between sperm and in vitro matured oocytes across early embryonic development stages. (h) Box plots of DNA methylation levels of in vivo derived MII oocyte-specific DMRs between sperm and in vitro matured oocytes across early embryonic development stages. (i) Box plots of DNA methylation levels of sperm-specific DMRs between sperm and GV oocytes across early embryonic development stages. (j) Box plots of DNA methylation levels of in vitro matured oocyte-specific DMRs in comparisons between sperm and in vitro matured oocytes across early embryonic development stages. (k) Heatmap of the methylation level of sperm-specific DMRs in comparisons between sperm and in vitro matured oocytes among different genomic regions across different developmental stages. **(l)** Heatmap of the methylation level of in vitro MII oocyte-specific DMRs in comparisons between sperm and in vitro matured oocytes among different genomic regions across different developmental stages. In each of these panels, the color keys from green to red indicate methylation levels from low to high.

Supplementary Figure S2.5



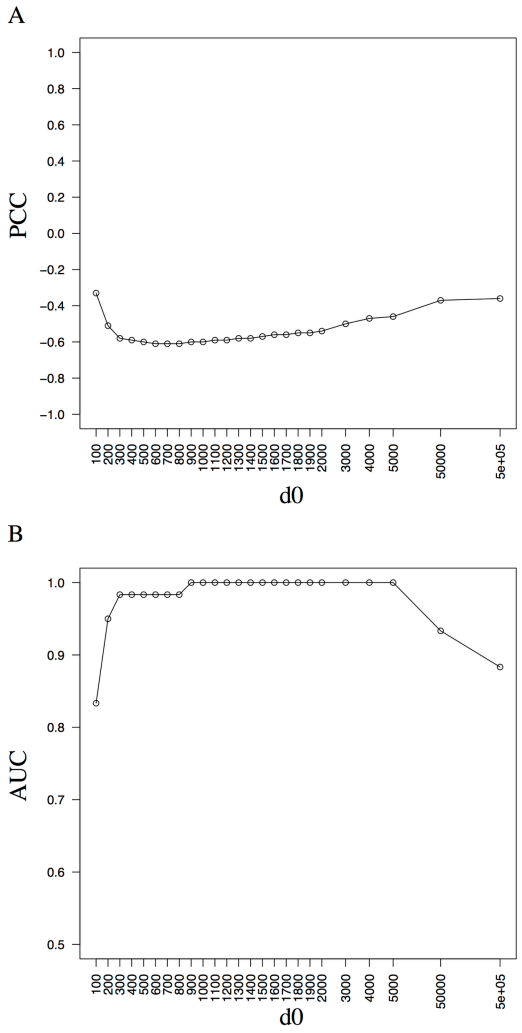
(a) Dynamics of DNA methylation of LINEs (L1, L2, and BovB) during bovine embryo development.
(b) Expression patterns of LINEs (L1, L2, and BovB) during bovine embryo development.

Supplementary Figure S3.1



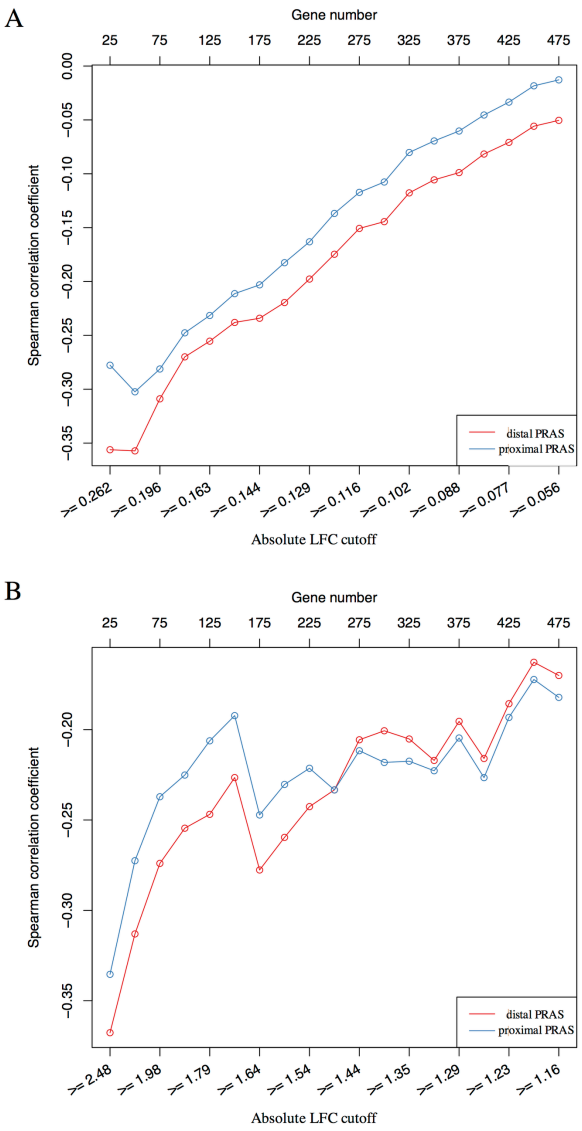
CELF4 binding characteristics in 3'UTRs. Shown are distributions of the distances between the iCLIP reads and the proximal/distal end of 3'UTRs in mRNAs. X-axis represents the distance (number of nucleotide) to the proximal/distal end of 3'UTRs. Y-axis represents the average iCLIP read counts within the significant peaks at that position across all the genes. The curve for the distal end is highlighted by red color and that for the proximal end is highlighted by blue.

Supplementary Figure S3.2



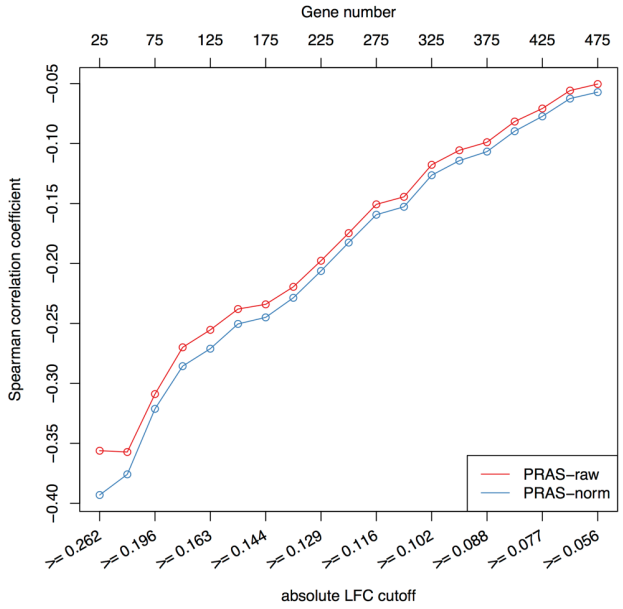
Correlation coefficient curve and AUC curve of PRAS with different d0s. (A) The line chart of Pearson's correlation coefficient between the gene score and the gene expression LFC in the qPCR-validated targets of CELF4. The X-axis represents the different d0s applied to PRAS and the Y-axis shows the value of Pearson's correlation coefficient. Each dot in the plot is for one d0 usage in PRAS. (B) Similar to A, but for the AUC values of the ROC analysis. These two line-charts show that the performance of PRAS is stable with the reasonable d0 selection around 1,000 nt.

Supplementary Figure S3.3



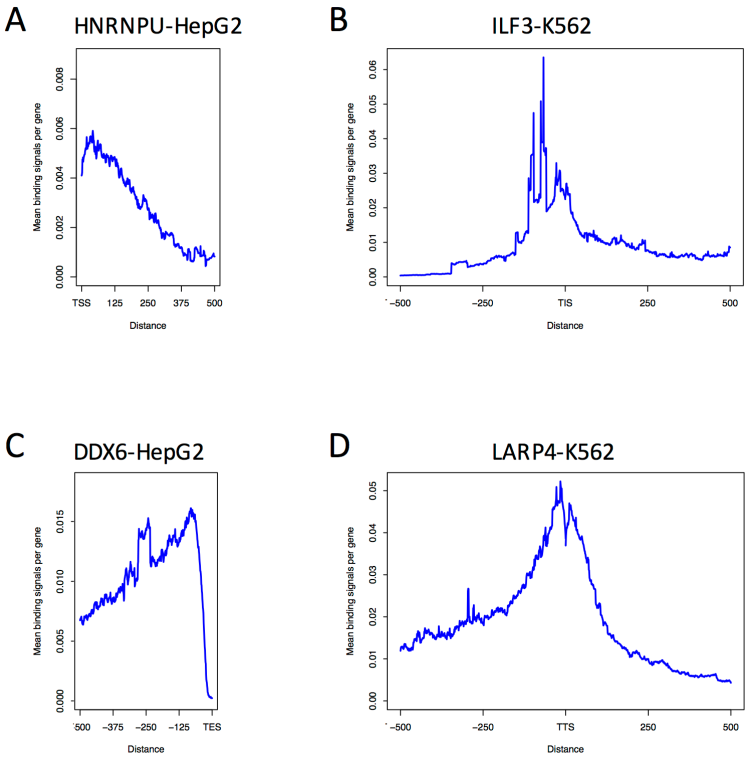
Correlation analysis of PRAS with different reference sites. (A) The line chart of Spearman's correlation coefficient between the gene score and the gene expression LFC in the Celf4-regulated list. The X-axis represents the different cutoffs applied to extract the subset of genes and the Y-axis shows the value of Spearman's correlation coefficient. The corresponding curves for distal PRAS and proximal PRAS are indicated by red and blue lines, respectively. Each dot in the plot is for one subset of genes selected based on the absolute LFC cutoff. (B) Similar to A, but for the Celf1-regulated list. These two line-charts show that the top ranked targets by distal PRAS have higher enrichment in the regulated lists comparing to those of proximal PRAS.

Supplementary Figure S3.4



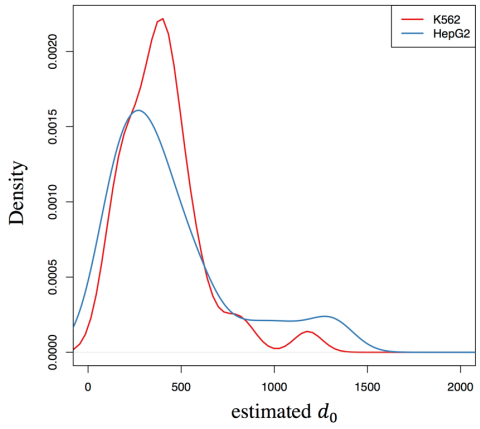
Correlation analysis of PRAS with different peak intensity input. The line chart of Spearman's correlation coefficient between the gene score and the gene expression LFC in the Celf4-regulated list. The X-axis represents the different cutoffs applied to extract the subset of genes and the Y-axis shows the value of Spearman's correlation coefficient. The corresponding curves for PRAS-raw and PRAS-norm are indicated by red and blue lines, respectively. Each dot in the plot is for one subset of genes selected based on the absolute LFC cutoff.

Supplementary Figure S3.5



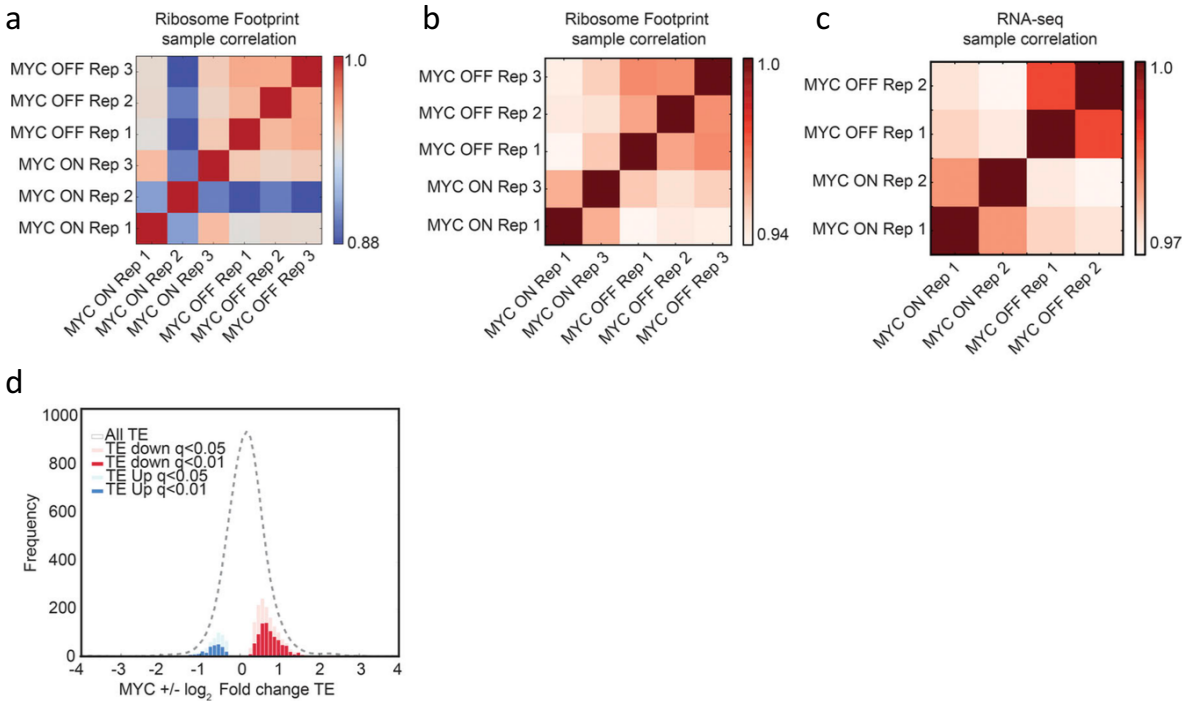
RBP examples of eCLIP signal distribution around different reference sites. (A) Shown are distributions of the distances between the HNRNPU eCLIP peaks and the transcription start site (TSS) in the mRNAs of the HepG2 cell line. X-axis represents the distance (number of nucleotide) to the TSS. Y-axis represents the average eCLIP enrichment ratio within the significant peaks at that position across all the genes. (B) Similar to A, but around the translation initiation site (TIS) for RBP ILF3 in K562 cell line. (C) Similar to A, but around the translation termination site (TTS) for RBP DDX6 in HepG2 cell line. (D) Similar to A, but around the transcription end site (TES) for RBP LARP4 in K562 cell line.

Supplementary Figure S3.6



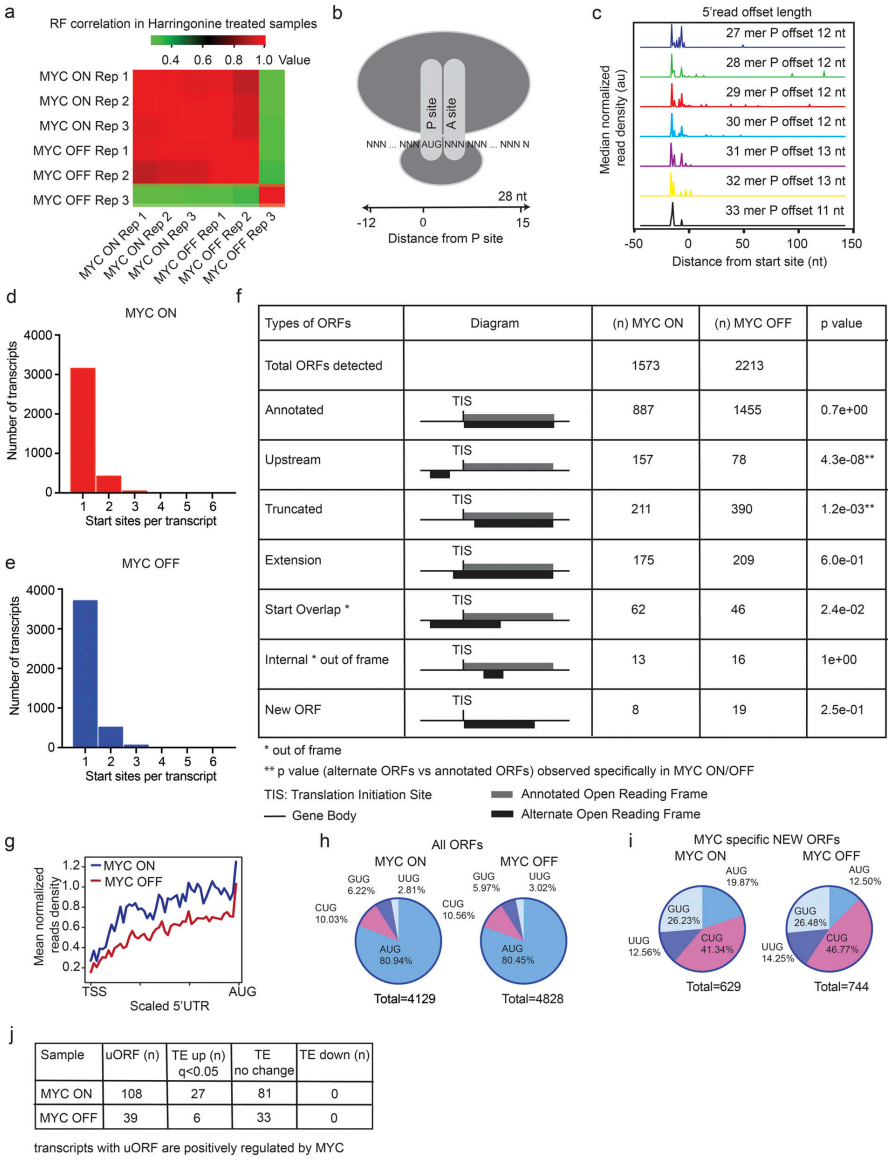
Distribution of the estimated decay parameter for PRAS. Shown are the distributions of the estimated d0 for PRAS in K562 and HepG2 cell lines. The density curves are highlighted by red and blue for RBPs in K562 and HepG2, respectively. The estimation is done based on the eCLIP peak intensities around the selected reference sites as described in the subsection “PRAS score is a strong predictor of PCR-validated mRNA targets of CELF4”.

Supplementary Figure S4.1



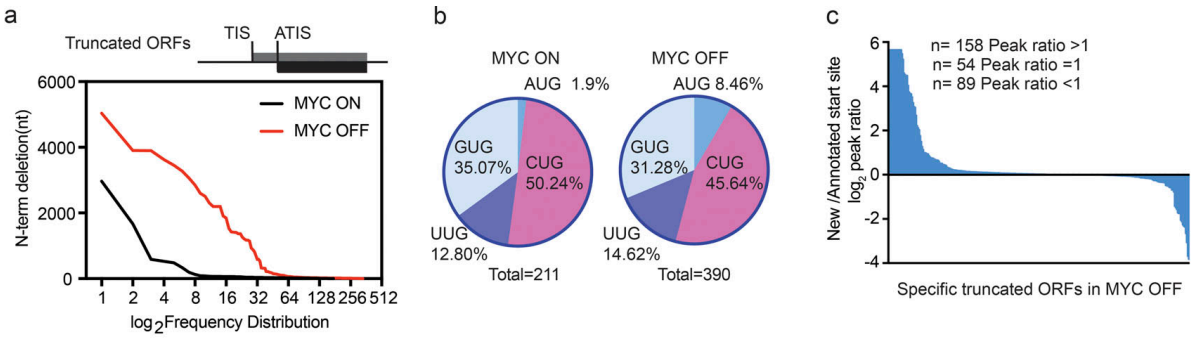
(a–c) Read count correlation plots of replicates from untreated and doxycycline (0.1 µg/ml)- treated ribosome footprinting and total RNA samples. (d) Frequency distribution of the ratio of TE in untreated (MYC ON) and doxycycline-treated (MYC OFF) P493-6 cells (TEMYC+/TEMYC−). TE up (red) indicates mRNAs that require MYC for translation; TE down (blue) are MYC independent. Biological replicates MYC ON: n = 2; MYC OFF: n = 3.

Supplementary Figure S4.2



MYC affects TIS choice. (a) Ribosome footprint read count correlation plots for indicated samples. (b) Schematic showing P-site offset for ribosome footprint reads relative to the TIS. (c) P-site offset measured for varying read length from 27 to 33 mer in ribosome footprinting samples. (d and e) Histogram of number of TISs per transcript in harringtonine-treated MYC ON (n = 3) and OFF (n = 3) samples. (f) ORFs detected specifically in MYC ON (n = 3) and OFF (n = 3) samples based on TIS detection in harringtonine-treated samples. (g) Metagene analysis of ribosome density per triplet across the 5'UTR of transcripts that show >50 reads per CDS in MYC ON (n = 3) and OFF (n = 3) samples treated with harringtonine. (h and i) Distribution of AUG codons and nearcognate NTG codons at predicted TISs in all ORFs and specific ORFs detected in MYC ON (n = 3) and OFF (n = 3) samples; significance by Fisher's exact test. (j) uORF-containing genes are associated with increased or unchanged, and not with decreased, TE.

Supplementary Figure S4.3



Low MYC expression favors downstream start sites that lead to functional N-terminal truncations. (a) Distance of ATIS detected from the annotated TIS for truncated ORFs in MYC ON (n = 3) and OFF (n = 3) samples. (b) Distribution of AUG codons and near-cognate NTG codons at predicted TISs for truncated ORFs in MYC ON (n = 3) and OFF (n = 3) samples; significance by Fisher's exact test. (c) RF peak height ratios of TISs at the ATIS versus the annotated TIS indicates relative usage of aberrant and annotated TIS for truncated ORFs in MYC OFF (n = 3) samples.

APPENDIX B: SUPPLEMENTARY TABLES

Supplementary Table S1.1. Sequencing qualities, read mapping, the average covered CpG sites at 1X, 5X, and 10X and the bisulfite conversion rate at each stage of bovine embryo development

Stages	No. of Total Sequencing Reads	No. of Mappable Reads	Mapping Ratio	Total Unique CpG Sites(1X) across stage	Total Unique CpG Sites(5X) across stage	Total Unique CpG Sites(10X) across stage	Bisulfite Conversion Rate
Sperm_1	42,995,939	8,171,800	19.05%	9,925,201	5,269,264	3,444,269	99.64%
Sperm_2	44,304,015	8,430,731	19.00%				99.64%
Sperm_3	44,790,334	8,644,057	19.30%				99.50%
GV oocyte_1	27,548,457	7,609,130	27.55%	3,329,486	2,427,043	2,149,315	99.64%
GV oocyte_2	62,635,569	22,494,131	36.00%				99.66%
GV oocyte_3	26,145,253	8,483,532	32.45%				99.62%
In vitro_Mature oocyte_1	69,330,722	31,984,613	46.20%	1,852,930	1,269,660	1,166,280	99.64%
In vitro_Mature oocyte_2	77,437,820	39,289,823	50.70%				99.66%
In vitro_Mature oocyte_3	21,010,326	14,453,744	32.35%				99.62%
In vivo_Mature oocyte_1	32,778,751	8,572,014	26.15%	3,974,950	2,806,316	2,100,123	99.60%
In vivo_Mature oocyte_2 *	25,802,891	657,022	2.55%				99.63%
In vivo_Mature oocyte_3	31,893,887	10,942,544	34.35%				99.68%
2cell_1	53,818,711	19,307,395	35.85%	4,990,351	3,827,203	3,370,405	99.62%
2cell_2	49,443,727	16,596,961	33.55%				99.59%
2cell_3	60,438,339	20,453,069	33.85%				99.59%

4cell_1	52,114,762	17,562,786	33.65%	3,351,058	2,608,784	2,362,163	99.61%
4cell_2	60,319,699	16,742,038	27.75%				99.62%
4cell_3 *	46,662,876	1,353,162	2.90%				99.26%
8cell_3	71,991,748	19,325,325	26.85%	11,621,580	6,872,748	4,883,138	99.61%
8cell_2	42,391,007	11,489,874	27.05%				99.56%
8cell_1	37,571,406	10,288,617	27.60%				99.53%
Early Morula_1	66,234,708	23,780,966	35.95%	4,879,048	3,194,842	2,506,589	99.60%
Early Morula_2	57,351,370	29,954,866	52.25%				99.54%
Early Morula_3	34,227,426	10,301,685	30.25%				99.66%
Compact Morula_1	22,604,321	7,124,383	31.80%	3,065,791	2,271,435	1,905,354	99.65%
Compact Morula_2	28,728,038	8,241,313	28.80%				99.60%
Compact Morula_3	32,199,795	13,910,721	43.45%				99.65%
Blastocyst_1	35,132,031	8,543,949	24.60%	5,810,378	3,527,677	2,510,965	99.69%
Blastocyst_2	38,343,174	10,486,860	27.45%				99.50%
Blastocyst_3 *	23,502,236	14,044	0.10%				99.62%

* sample removed for downstream analysis.

Supplementary Table S1.2. List of top 20 genes that promoter methylation was significantly and inversely correlated with gene expression at the 8-cell stage.

gene	meth	log2(8cell_FPKM)
RPL32	0.004385963	12.49091855
RPS3	0.006448877	13.22399414
RPL37	0.009624113	11.99353639
HMGN2	0.005339773	9.503799904
RPLP1	0.01110277	11.18097828
RPS17	0.011714179	11.62215709
IMP3	0.005382453	8.787687118
NDUFB1	0.004937026	8.713455254
KPNA2	0.005543985	8.785423042
TPT1	0.012550528	11.97897495
UQCRFS1	0.010416674	9.354668133
RPS28	0.012918541	11.68779557
AURKA	0	8.057628973
RPL36AL	0.01083076	8.905573535
SKP1	0.010741267	8.634622308
PPIA	0.013904638	11.81528722
UQCRB	0.007638385	7.937544491
RAB1A	0.005740744	7.759828278
OAZ1	0.013148499	9.809655602
NDUFB8	0.013184748	9.86500014

Supplementary Table S1.3. The hypergeometric enrichment analysis of the hypermethylated and hypomethylated tiles in bovine gametes, exhibited the strong enrichment for different genomic regions (hypergometric enrichment test).

Genomic regions	Sperm vs. In vivo MII (P value)		Sperm vs. In vitro MII (P value)		Sperm vs. GV (P value)	
	Hypermethylated	Hypomethylated	Hypermethylated	Hypomethylated	Hypermethylated	Hypomethylated
Exons	1.90E-125	1.00E-147	3.50E-55	6.90E-37	1.20E-34	1.50E-80
Introns	2.30E-182	1	1.40E-58	1	6.90E-48	1
Promoters	1	0	1	0	1	0
LINE	0.0037	1	0.014	1	0.12	1
SINE	0.00055	1	0.12	1	0.11	1
LTR	1	1	1	1	1	1
CGI	1	0	1	0	1	0

Supplementary Table S1.4. Methylation levels of DMRs in in vitro_MII and in vivo_MII oocytes. Only tiles that are highly methylated in in vivo_MII oocytes are shown.

Tile	In vivo_MII	In vitro_MII	Genomic region	Gene body
chr24_27695	50.0%	49.6%	NA	NA
chr29_493730	100.0%	97.8%	NA	NA
chr7_166679	100.0%	83.6%	exon	SMARCA4
chr4_53396	100.0%	79.7%	intron	DDC
chr22_611969	100.0%	71.2%	NA	NA
chr28_265967	50.0%	33.3%	SINE	PPA1
chr28_436501	100.0%	65.9%	intron	WDFY4
chr25_358525	50.0%	28.9%	SINE	COL26A1
chr4_1146632	50.0%	20.4%	LINE	NA
chr13_774167	100.0%	33.3%	NA	NA
chr21_597250	100.0%	33.3%	NA	NA
chrUn_GJ057830v1_1	100.0%	14.5%	NA	NA
chr1_52674	50.0%	7.1%	NA	NA
chrUn_GJ058425v1_1437	100.0%	4.8%	NA	NA
chr1_101969	100.0%	0.0%	LINE	NA
chr1_671281	100.0%	0.0%	LINE	NA
chr12_129879	100.0%	0.0%	NA	NA
chr15_849050	100.0%	0.0%	NA	NA
chr16_580319	100.0%	0.0%	LINE	TNR
chr16_784219	50.0%	0.0%	NA	NA
chr19_360207	100.0%	0.0%	NA	NA
chr19_580299	50.0%	0.0%	NA	NA
chr20_697797	100.0%	0.0%	NA	NA
chr20_72106	100.0%	0.0%	LINE	NA
chr22_109572	100.0%	0.0%	intron	ITGA9
chr25_329553	100.0%	0.0%	SINE	NA
chr26_452533	100.0%	0.0%	NA	NA
chr5_1192280	50.0%	0.0%	NA	NA
chr7_1054652	100.0%	0.0%	SINE	NA
chr9_876387	100.0%	0.0%	LINE	NA
chrX_12476	50.0%	0.0%	NA	NA
chrX_544118	100.0%	0.0%	NA	NA
chrX_566496	50.0%	0.0%	LINE	NA
chrX_725651	100.0%	0.0%	LTR	NA